AD______________

Award Number: DAMD17-03-1-0015

TITLE: A Likelihood Ratio Classifier for Computer-aided Diagnosis in Mammography

PRINCIPAL INVESTIGATOR: Anna O. Bilska-Wolak
Carey E. Floyd, Jr., Ph.D.

CONTRACTING ORGANIZATION: Duke University
Durham, North Carolina 27710

REPORT DATE: July 2004

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** July 2004	**3. REPORT TYPE AND DATES COVERED** Annual Summary (13 Jun 2003 - 12 Jun 2004)

4. TITLE AND SUBTITLE
A Likelihood Ratio Classifier for Computer-aided Diagnosis in Mammography

5. FUNDING NUMBERS
DAMD17-03-1-0015

6. AUTHOR(S)
Anna O. Bilska-Wolak
Carey E. Floyd, Jr., Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University
Durham, North Carolina 27710

E-Mail: aob@duke.eau

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. Abstract ***(Maximum 200 Words)*** *(abstract should contain no proprietary or confidential information)*

Although screening x-ray mammography has become a very sensitive method for detecting breast cancer, mammography has low specificity in its diagnostic stage. About 67-85% of breast biopsies are performed on benign lesions. Because of cost and detrimental effects of unnecessary biopsies, the number of biopsies performed on benign lesions needs to be reduced. In this research we are developing a highly sensitive and specific computer-aided diagnosis classifier based on the likelihood ratio, which is designed to aid physicians to identify lesions that should not be sent to biopsy. The classifier is being developed using a large database of over four thousand breast biopsy cases from several medical centers. The cases present in the databases are described using BI-RADSTM lexicon and patient history. The product of this year's research includes a first generation likelihood ratio classifier that verifies the initial hypothesis. The performance of the classifier was comparable to or better than other classifiers previously developed for breast biopsy classification. The preliminary results suggest that the LR is a robust classifier for prediction of biopsy outcome. By decreasing the number of benign cases sent to biopsy, the classifier could be a valuable tool for physicians and ultimately beneficial to hospitals and patients.

14. SUBJECT TERMS
computer-aided diagnosis, mammography, likelihood ratio, biopsy, case-based reasoning, artificial intelligence

15. NUMBER OF PAGES
64

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** Unlimited

Table of Contents

INTRODUCTION

Although screening x-ray mammography has become a very sensitive method for detecting breast cancer, mammography has low specificity in its diagnostic stage. About 67-85% of breast biopsies are performed on benign lesions. Because of cost and detrimental effects of unnecessary biopsies, the number of biopsies performed on benign lesions needs to be reduced. In this research we are developing a highly sensitive and specific compute-aided diagnosis classifier based on the likelihood ratio, which is designed to aid physicians to identify lesions that should not be sent to biopsy. The classifier is being developed using a large database of over four thousand breast biopsy cases from several medical centers. The cases present in the databases are described using BI-RADS^{TM} lexicon and patient history, and represent the collective knowledge of physicians. The resulting classifier will be statistically based, mathematically simple, and computationally efficient. Rigorous and exhaustive classifier evaluation methods include Receiver Operating Characteristic (ROC) analysis and leave-one-out bootstrap sampling.

STATEMENT OF WORK (01-2004)

Task 1. Develop and optimize case representation and database of over 4500 biopsy cases. (Months 1-36)

a. Previously acquired cases from Duke University, University of Pennsylvania (Penn)

b. Continue extracting information for cases from the (DDSM) database of University of South Florida

c. Acquire cases from other medical institutions

Task 2. Develop the LR and optimize its subcomponents. (Months 1-24)

a. Optimize mathematical feature representation from categorical case data

b. Estimate and optimize the N-dimensional density distribution of features (histogram approach, histogramming with smoothing functions, nearest-neighbor approaches, optimal decision fusion, kernel-density estimation)

c. Optimize features used (exhaustive search techniques, singular value decomposition, principle component analysis)

d. Evaluate model using ROC analysis, Round Robin sampling, and bootstrap

Task 3. Evaluate the performance of the LR under various conditions stemming from the input data. (Months 12-30)

a. Train and test separately on data from different institutions (i.e. train on all cases, test on cases from Duke University)

b. Train and test separately on different lesion types (i.e. train on all cases, test on mass lesions)

Task 4. Simulate and evaluate the use of LR in a clinical setting. (Months 24-36)

a. Analyze the optimized classifier on a set of data not used in training/development.

b. Examine how the standards (sensitivity, specificity) set on the training data affect the sensitivity and specificity on the new test data

c. Establish guidelines for retraining the classifier when a significant amount of new data is added

d. Conduct a retrospective clinical evaluation to evaluate LR's influence on physician's performance.

BODY

Task 1 is an ongoing effort to collect more cases. A large database of cases has been obtained and adapted for the project. The initial database has been increased by approximately 400 cases from Duke University, which were previously unavailable to us. Approximately 400 cases from Sloan-Kettering Cancer Institute, 600 cases from the University of North Carolina, and 125 cases from University of Maryland have also been recently obtained. Case acquisition and selection will continue as possible, although this task has been completed and outdone as presented in the Statement of Work.

As specified in **Task 2**, a first generation likelihood ratio-based (LRb) classifier has been developed. The likelihood ratio is the optimal detector to determine the presence or absence of a signal in noise.[1-5] Given features describing the presence of a suspicious lesion on a mammogram, one has to decide whether or not a malignancy - a signal - is present. The likelihood ratio λ is optimal under the assumption that full knowledge of the statistical properties of the data is known. The likelihood ratio is also referred to as the ideal observer.

The hypothesis that the signal is present is the H1 hypothesis, and the "no signal" or null hypothesis is referred to as H0. The decision whether the signal is present or absent is optimized by thresholding the likelihood ratio $\lambda(V)$,

$$\lambda(V) = \frac{p(V \mid H1)}{p(V \mid H0)}$$

where the $p(V|H1)$ represents the probability density function (PDF) of the available data V under the "signal present" hypothesis, and $p(V|H0)$ represents the PDF of the available data under the "no signal" hypothesis. The signal to be detected is the malignancy of the breast lesion. Therefore, we used the available descriptions for malignant biopsy cases to represent the H1 distribution, and the descriptions for benign biopsy cases to represent the H0 distribution. More detail on classifier development is also available.[6, 7]

Task 2a. Specifically, the biopsy cases present in our database were described with BI-RADS™ lexicon, which is a collection of categorical descriptives for features (Table 1). We evaluated two approaches: 1) ranking and 2) histogram-based. 1) The categorical descriptives for each feature were arranged in order of increasing risk of malignancy, and a ranking scale was assigned. This scale had been established in consultation with physicians. This ranking scale was used in conjunction with nearest neighbor approaches (see Task 2b). 2) While the ranking approach is practical and has resulted in commendable performance, it is possible that better performance can be achieved without the use of a classifier that depends on a numerical or ranking scale. This is true because while one finding might be considered more malignant than another, real-life performance and database content might not be reflected in the ranking assignments. For example, physicians might consider dystrophic calcifications more at risk to be malignant than round, and thus designate a rank of 10 to dystrophic calcifications, and 9 to round calcifications. However, physicians might assign round more often to malignant calcifications than dystrophic, thus behaving in opposition to their scale. Any classifier that utilizes averaging or interpolation of findings will be dependent on this scale/ordering of the categorical features, and may be affected unfavorably. For the second pass, therefore, we have chosen to represent the categorical features as discrete histogram distributions. Such nonparametric models can be very effective, and also reduce the risk of misinterpreting the data.[8] Conversely, parametric models often interpolate the data, yet it might be unnatural to average categorical feature findings.

We have concluded that the best feature representation is one that is independent of ranking scales, and follows naturally from the data presented - the histogram approach. Our data supports this conclusion, since best performance so far has been achieved with the histogram-encoded version of the classifier.[7]

Table 1: BI-RADS^{TM} feature representation using the ranking approach. The ranking was established in consultation with mammographers.

Feature	Finding	
(1) Patient Age	years	
(2) Mass Margin	no mass	0
	well-circumscribed	1
	microlobulated	2
	obscured	3
	ill-defined	3
	spiculated	4
	no info	-1
(3) Mass Density	no mass	0
	fat-containing	1
	low-density	2
	isodense	3
	high-density	4
	no info	-1
(4) Mass Shape	no mass	0
	round	1
	oval	2
	lobular	3
	irregular	4
	no info	-1
(5) Mass Size	mm	
(6) Menopause	pre-menopausal	1
	post-menopausal	2
	no info	-1
(7) Calcification Distribution	no calcs	0
	diffuse	1
	regional	2
	segmental	3
	linear	4
	clustered	5
	no info	-1
(8) Calcification Morphology	no calcs	0
	milk of calcium-like	1
	eggshell or rim	2
	skin	3
	vascular	4
	spherical or lucent-centered	5
	suture	6
	coarse ("popcorn")	7
	large rod-like	8
	round	9
	dystrophic	10
	punctate	11
	indistinct	12
	pleomorphic	13
	fine branching	14
	no info	-1
(9) Calcification Number	no calcs	0
	<5	1
	5 to 10	2
	>10	3
	no info	-1
(10) Associated Findings	none	0
	skin lesion	1
	hematoma	2
	post surgical scar	3
	trabecular thickening	4
	skin thickening	5
	skin retraction	6
	nipple retraction	7
	axillary adenopathy	8
	architectural distortion	9
	no info	-1
(11) Special Cases	none	0
	intrammamary lymph node	1
	assymetric breast tissue	2
	focal assymetric density	3
	tubular density or solitary dilated duct	4
	no info	-1
(12) Quad	posterior	1
	central	2
	LIQ	3
	LOQ	4
	UIQ	5
	UOQ	6
	axillary tail	7
	subareolar	8
	no info	-1
(13) Change from prior	no change	0
	new lesion	1
	qualitative change	2
	quantitative change	3
	no info	-1
(14) Architectural Distortion as main finding	none	0
	present	1
	no info	-1
(15) Hormone Use	no hormone use	1
	hormones (such as BCP estrogen, progesterone)	2
	no info	-1
(16) Breast Side	left	1
	right	2
	no info	-1

The next challenge (**Task 2b**) concerned the logistics of the representation of the H0 and H1 feature distributions for the biopsy cases. While it would be ideal to compute the f-variate distribution of all the features for the LRb, populating the 16-dimensional feature space presents a problem: an extremely large number of cases would be needed to provide an adequate representation of the 16-dimensional space. There are possible solutions to this problem that we have utilized: 1) estimating the distributions with nearest-neighbor approaches, 2) with histogram approach.

Task 2b. Nearest Neighbor approach.
In the nearest neighbor approach, we compared a test case to a reference (training) collection of cases, and identified similar cases. Euclidean distance measure was used to determine the similarity between the test case and the reference cases. Each finding was first normalized using linear scaling to unit range. The distance between a test case and a reference case is thus,

$$D_{\mathrm{Euclidean}}(\mathrm{test,ref}) = \sqrt{\sum_{i=1}^{n} |\phi_{i_{\mathrm{test}}} - \phi_{i_{\mathrm{ref}}}|^2},$$

where n is the number of features (2 in this setup), and ϕ is the normalized feature value for the specific case (test or reference), and feature i (mass margin or age). Given this distance between the test case and a reference case, the two cases were judged to be similar if the distance between them was less than a specified similarity threshold. In an event where the test case matches none of the reference cases, the test case is automatically assigned a decision variable of 1, resulting in a malignant classification for the test case. The best similarity threshold was obtained by exhaustively examining all possible thresholds and maximizing partial ROC area (0.90AUC). The nearest neighbor approach of finding similar cases allows us, in effect, to create running averages of the feature distributions. The exact process of averaging depends on the distance measure of choice. Results of this approach and the exhaustive search for the best similarity distance and feature combination is presented below in 2c, feature optimization.

Histogram approach of optimizing density. Given the small number of findings for each of the BI-RADS™ features, we can represent the densities as histograms with fixed bins. This will allow the findings of a feature to remain separate and unaffected by ordering. For example, since mass shape has only four possible descriptions, four bins will be designated for each finding. We have also introduced a fifth bin, to represent the lack of information, or "no info." Therefore, the mass shape histogram has five possible bins. A different strategy was chosen for the continuous findings, such as age and mass size. Creating a histogram for these is slightly more complex due to the large number of possible values. We used Scott's rule[8-10] to determine the optimal histogram bin width h. Let σ represent the standard deviation and n the number of available observations. The optimal bin width h is then,

$$h=3.5*\sigma*n^{-1/3}$$

The interval within two standard deviations of the mean was subdivided by the bin width. Observations falling outside the two standard deviations were included in extreme right and left bins. For the mass size distribution, this resulted in nine bins. An extra bin indicating "no information" was also added, resulting in ten bins for the mass size distribution. Essentially, in this setup the encoding of categorical (BI-RADS™) findings matters little. The only fact of importance is that each finding remains a separate category. We eventually encode the findings as numbers in the algorithm, because it is easier for a computer program to use them, but the values we choose can be any number. More information on density estimation using histogram approach is available.[7]

The optimization of the density distribution will continue. In summary, we have performed nearest-neighbor approaches and histogram representation.[7] Little potential is now foreseen for the wavelet and frequency approaches, and in alternative we will work on kernel density estimation.

For feature optimization (**Task 2c**), singular value decomposition and principle component analysis remain to be evaluated, while for nearest neighbor approach exhaustive search technique has been completed.

Task 2c. All possible combinations of ten features were examined to optimize the similarity selection criteria on 1027 mammographic cases from Duke. Two distance measures, Euclidean and Hamming, were used for the nearest-neighbor approach. In all, 1023 feature combinations (strategies) were investigated. Out of the 1023 strategies examined, the strategy that produced the largest ROC area (0.818 ± 0.013) for the Hamming distance measure included comparison of six of the features: calcification distribution, calcification number, calcification morphology, mass margin, mass shape, and age. The strategy that produced the largest ROC area (0.822 ± 0.013) for the Euclidean distance measure included six of the features: calcification number, calcification morphology, mass margin, mass shape, special findings, and age. The results of the ROC analysis for 200 bootstrap samples of the top strategies for each distance measure are summarized in Table 2, showing mean values and standard deviations. Based on these results, the most influential features appear to be patient age, mass margin, mass shape, and calcification morphology.

Table 2: Performance summary for top feature combinations of the Hamming and Euclidean distance measures.

Distance Measure	AUC	0.90AUC	False Positive Fraction at 100% Sensitivity	False Positive Fraction at 98% Sensitivity	False Positive Fraction at 95% Sensitivity
Hamming	0.818 ± 0.013	0.393 ± 0.028	0.796 ± 0.044	0.664 ± 0.023	0.608 ± 0.036
Euclidean	0.822 ± 0.013	0.423 ± 0.030	0.913 ± 0.109	0.684 ± 0.068	0.537 ± 0.033

As specified in **Task 2d**, The performance of the classifier was evaluated using Receiver Operating Characteristic (ROC) analysis. We were most interested in the performance at high sensitivities, because missing a malignancy is a graver risk to the patient. Therefore, the main performance metrics throughout this research included: (1) area under the curve (AUC), which represents average sensitivity over all specificities, (2) partial area index (PAI) above the sensitivity of 90%, which represents the performance of classifier at high sensitivities, and (3) the specificity at 98% sensitivity. The last metric helps indicate how many benign cases could be potentially spared from biopsy while correctly diagnosing 98% of malignancies. This means that 2% of the malignancies could have potentially delayed biopsy and treatment. Since the cases in our database are diagnostic mammography cases, and the physician is already aware of the suspicious lesion, it is conceivable that the performance of the classifier with the physician could be 100% sensitivity.

In summary, ROC and the other evaluation techniques are under consistent utilization for testing each new modification to the classifier, as can be seen throughout this BODY and references.[6, 7, 11]

Task 3. The acquisition of new cases from other medical institutions substantially increases the effort required to complete this task. An evaluation has been carried out on a) a subset of cases from two institutions, Duke University and University of Pennsylvania, and b) on different lesions types from Duke University data.

Task 3a. Evaluation on different lesions from Duke University.

The database used consisted of 1433 biopsy-proven cases from Duke University Medical Center. The cases were divided into 3 categories: mass cases (646 total, 233 malignant), calcification cases (653,219), and other (134,50). Each mass case contained a mass lesion, and possibly also contained calcifications and other findings. Calcification cases contained calcifications and possibly other findings, but no mass lesions. Other cases included cases with all other findings, but did not contain masses or calcifications. Patient age varied from 24 to 89, with a mean age of 55 years.

The performance of the classifier on each case type is presented in Table 3. The classifier performed best on the mass cases, resulting in a highest $_{0.90}$AUC of 0.60, and an AUC of 0.92. The feature combination resulting in the highest $_{0.90}$AUC for mass cases included four of the features: mass margin, calcification morphology, associated findings, and age. At 98% sensitivity, the classifier would correctly spare 209 (51%) benign mass lesions while misclassifying 5 (2%) malignancies. The Round Robin ROC curve for mass cases is presented in Figure 1A, while the partial ROC curve (above 90% sensitivity) is shown in Figure 1B. Note that the values presented in Table 3 are mean values from the bootstrap evaluation, and not direct values from the Round Robin curve.

Table 3: Performance summary of LRb on each lesion type from Duke University.

Case Type	A_Z ± STD	$_{0.90}A_Z$ ± STD	TPF at FPF=0.95	Number of benign cases spared at FPF=0.95	TPF at FPF=0.98	Number of benign cases spared at FPF=0.98
Masses	0.92 ± 0.01	0.60 ± 0.05	0.34 ± 0.05	274 (66%)	0.49 ± 0.13	209 (51%)
Calcifications	0.64 ± 0.02	0.16 ± 0.03	0.83 ± 0.03	73 (17%)	0.88 ± 0.03	50 (12%)
Other	0.76 ± 0.04	0.33 ± 0.10	0.64 ± 0.12	30 (36%)	0.78 ± 0.13	18 (22%)

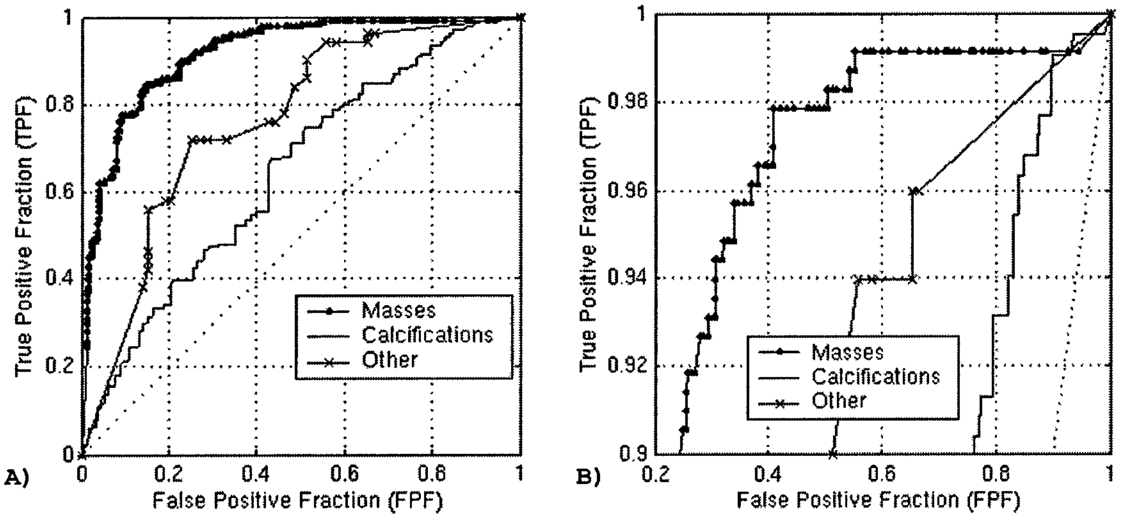

Figure 1: A) ROC curves for masses, calcifications, and other cases. B)Partial ROC curves for masses, calcifications, and other cases.

The classifier performed substantially better on the mass cases than on the calcifications. The AUC for the mass cases was 0.92, while the calcification cases had an AUC of only 0.64. The difference in performance is also very evident in the high sensitivity region of the ROC curve. At 98% sensitivity, the classifier would spare 51% benign mass cases, while only 12% benign calcifications would be spared at this sensitivity.

Task 3b Evaluation on cases from Duke University and University of Pennsylvania.[11]

The data set of mammographic mass cases from the Duke University Medical Center consisted of 645 biopsy cases. Of these, 232 were malignant and 413 proved benign at biopsy. Another data set from the University of Pennsylvania Medical center consisted of 496 mammographic mass cases. In this subset, there were 296 malignant cases, and 200 benign. For this evaluation, each case in the database was represented by two features. One of the features was mass margin, described in accordance with the BI-RADS^{TM} lexicon. The description of mass margin was assigned by experienced mammographers at the time of the decision to biopsy. Mass margin can range from "well circumscribed" to "spiculated." The other feature used was patient age, which is a well known risk-factor for breast cancer.

In order to evaluate the performance of the classifier on the data from the two medical centers, we used four evaluation approaches. (A) The LRb was trained and tested on Duke data, (B) LRb was trained on Pennsylvania data, tested on Duke data, (C) LRb was trained and tested on Pennsylvania data, and (D) LRb was trained on Duke data, tested on Pennsylvania data.

Since several methods exist for training and testing on the same data, approach A was further separated into approaches A1 and An. In approach A1, a leave-one-out sampling approach was applied to evaluate the performance of the Duke-trained classifier on the Duke data set. Each case from the Duke set was used as a test case, while the remainder of the Duke cases was used as the training set. This was repeated over all cases, until all cases had been used as a test case. The ROC curve was produced by thresholding the likelihood ratio values of the test cases. In approach An, the classifier was trained and tested on all cases. This produced a consistency (resubstitution) curve for the Duke set. A similar procedure was followed for C, for training and testing on the Pennsylvania data set. The leave-one- out approach approach is referred to as C1, and the consistency approach is referred to as Cn. For approach D (cross-training across different institutions) all of the Duke data was used for training the classifier, and then all of the Pennsylvania data was used for testing. In a similar fashion for approach B, all of the Pennsylvania data was used for training the classifier, and then all of the Duke data was used for testing performance. In each situation, all of the training dataset was used to establish the threshold that would spare 98% of the benign cases from the training dataset. For approach D, the 98% sensitivity threshold from approach An was used. This means that the threshold was established using all of the Duke cases (the consistency curve). This threshold was then applied to the ROC curve of the test set. From this ROC curve, the new resulting sensitivity and specificity on the test were acquired. These results are presented in Table 4. Similarly for approach B, the 98% sensitivity threshold from approach Cn was used to establish the new sensitivity and specificity on the test set.

Testing the trained classifier on the Duke dataset (Approaches A-B) When the Duke dataset was used for training and testing the classifier in a leave-one-out fashion, the AUC was 0.91 ± 0.01 (Table 4, Figure 2). When the Penn dataset was used for training the classifier, the performance of the classifier tested on the Duke set was again 0.91 ± 0.01. While no difference was observed in the AUC, there was a small difference in the PAI, above the 90% sensitivity level. The PAI for testing and training on Duke was 0.59 ± 0.05, while it was only 0.53 ± 0.07 for training on Penn and testing on Duke. The PAI was lower for testing and training on data across different medical centers, while the AUC remained the same. The difference in PAI, was not statistically significant (p=0.57).

Table 4: Performance of classifier on Duke vs. Pennsylvania data.

Approach	Training Dataset	Testing Dataset	AUC	PAI (above 90% sensitivity)	# Benign Cases Potentially Spared at 98% Sensitivity
A_I	Duke	Duke (leave-one-out)	0.91 ± 0.01	0.59 ± 0.05	209 (51%)
A_{II}	Duke	Duke (consistency)	0.94 ± 0.01	0.71 ± 0.03	264 (64%)
B	Pennsylvania	Duke	0.91 ± 0.01	0.53 ± 0.07	see Table 2
C_I	Pennsylvania	Pennsylvania (leave-one-out)	0.85 ± 0.02	0.30 ± 0.07	28 (14%)
C_{II}	Penn	Penn (Consistency)	0.90 ± 0.01	0.52 ± 0.06	78 (39%)
D	Duke	Pennsylvania	0.85 ± 0.02	0.29 ± 0.07	see Table 2

Testing the trained classifier on the Pennsylvania dataset (Approaches C-D)

When the Pennsylvania dataset was used for training and testing the classifier, the leave-one-out AUC was 0.85 ± 0.02. When the classifier was trained on the Duke dataset, and tested on the Pennsylvania dataset, the AUC was also 0.85 ± 0.02. Therefore, the performance of the classifier tested on the Pennsylvania data was the same in terms of AUC, regardless of which dataset was used for training. While no difference was observed in the AUC, there was a small difference in the PAI. The PAI for testing and training on the Pennsylvania data was 0.30 ± 0.07, and 0.29 ± 0.07 for training on the Duke data. This difference was also not statistically significant (p=0.99).

Table 5: Performance of the cross-trained classifier using 98% sensitivity threshold established on the training data.

Approach	Training Dataset	Testing Dataset	Actual Resulting Sensitivity	# Benign Cases Potentially Spared from Biopsy
B	Pennsylvania	Duke	98%	189 (46%)
D	Duke	Pennsylvania	92%	104 (52%)

After training the classifier in a leave-one-out fashion on the Pennsylvania data, the LRb could potentially spare from biopsy 14% of the Pennsylvania benign lesions. Training on the Duke dataset and using the 98% threshold established on the Duke set would result in 92% sensitivity on the Pennsylvania data set. This sensitivity was also the next highest sensitivity point after 100% on the Pennsylvania curve. Using the classifier at this sensitivity would potentially obviate 46% benign cases from the Pennsylvania data set. These results suggest that while it is possible to train and test on data from different medical centers, it may be more beneficial to train and test on data from the same medical center.

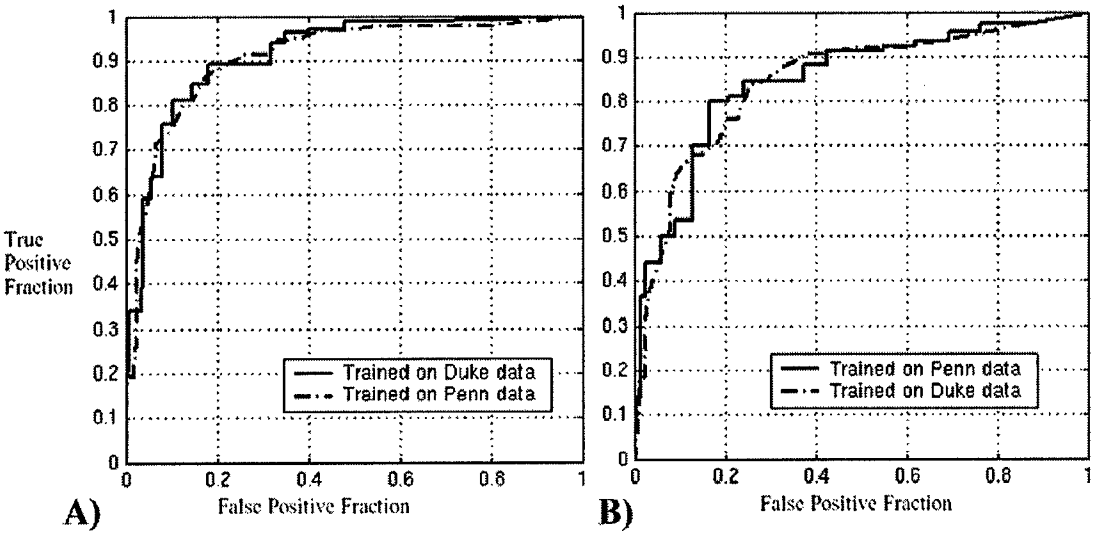

Figure 2: Performance of the classifier tested on on A) data from Duke University, and B) data from University of Pennsylvania.

It is difficult to ascertain why the performance of the classifier was lower on testing on the Pennsylvania data than on the Duke data. Given that the performance remains comparable regardless of which dataset is used for training, it is likely these databases do not contain more information for improvement of results. It is also difficult to ascertain why training on the Duke dataset and establishing the 98% sensitivity threshold did not produce equally high sensitivity in the testing on the Pennsylvania set. The threshold established on the Duke set resulted in a sensitivity drop to 92% on the Pennsylvania dataset. This sensitivity was also the next highest sensitivity point after 100% on the Pennsylvania curve. Since these results did not involve resampling of the inputs, more research is needed to examine why this drop in sensitivity was observed. However, it is encouraging that training the classifier on the Pennsylvania database did not result in a sensitivity drop. The 98% threshold established on the Pennsylvania set resulted in equally high sensitivity of 98% on the Duke set. The classifier would still be able to maintain 98% sensitivity on the Duke data at the training threshold, while potentially sparing from biopsy 46% of the benign cases.

While the performance of the classifier decreased at high sensitivities when training on a dataset from another medical center, the overall performance (AUC) remained the same regardless of which dataset was used for training. These results suggest that it may be possible to use data from multiple medical centers to create a global classifier for breast cancer prediction.

Since the LRB is being developed on a large multi-institutional database, a very comprehensive evaluation is still needed to evaluate national trends that could be translated into the clinic.

Task 4. An independent evaluation has been completed using a subset of the newly acquired cases. These included just mass cases from one institution.

For this evaluation,[12] we utilized the LRb developed on the database of 670 mass cases[7]. The 670 cases (245 malignant) from one medical institution were described using 16 features from the BIRADS™ lexicon and patient history findings. Continued data collection yielded additional 151 (43 malignant) cases that were previously unseen by the classifier. These new cases were examined by the developed classifier. Performance evaluation methods included Receiver Operating Characteristic (ROC), round-robin, and leave-one-out bootstrap sampling. The performance of the classifier on the training data yielded an average ROC area of 0.90+/- 0.02, and partial ROC area (0.90AUC) of 0.60+/-0.06. The exact non-parametric performance on the independent set of 151 cases yielded a ROC area of 0.88 and 0.90AUC of 0.57. Using a 100% sensitivity cutoff threshold established on the training data, the classifier was able to correctly identify 100% of the malignant lesions in the new independent set, while potentially obviating 26% of the biopsies performed on benign lesions. In this pre-clinical evaluation, the LRb classifier performed equally well on the new independent data that was not used for classifier development. The LRb classifier performance compared favorably with an artificial neural network. The LRb classifier shows promise as a potential aid in reducing the number of biopsies performed on benign lesions. More detail is available.[12] (Manuscript in submission, see Appendix).

In summary, this independent evaluation yielded promising results. Our continued research has shown that our earlier estimates on the effort required to complete the clinical evaluation were inadequate. A comprehensive clinical evaluation will be required to evaluate potential effects in the clinic given our new extensive data set from multiple institutions.

KEY RESEARCH ACCOMPLISHMENTS

Task 1. Develop and optimize case representation and database of over 4500 biopsy cases. (Months 1-36)

- The initial hypothesis has been verified experimentally.
- A large database of cases has been obtained and adapted for the project.
- The initial database has been increased by approximately 400 additional cases from Duke University, 400 cases from Sloan-Kettering Cancer Institute, 600 cases from the University of North Carolina, and 125 cases from University of Maryland.
- Although case acquisition will continue as possible, this task has been completed as proposed in the Statement of Work.

Task 2. Develop the LR and optimize its subcomponents. (Months 1-24)

- A first generation classifier has been created and implemented.
- The mathematical feature representation has been optimized for the current data and classifier.
- The N-dimensional density distribution of features has been optimized using the nearest-neighbor approach and histogram approach.
- Use of ROC analysis, Round Robin sampling and bootstrap are under consistent utilization for each classifier version.
- This task is 50% completed.

Task 3. Evaluate the performance of the LR under various conditions stemming from the input data. (Months 12-30)

- Training and testing of the classifier has been performed on cases from two medical centers.
- Training and testing of the classifier has been performed on different lesion types (masses vs. calcifications) from one institution.
- This task is 20% completed.

Task 4. Simulate and evaluate the use of LR in a clinical setting. (Months 24-36)

- An independent evaluation on a set of mass cases previously unseen by the classifier has been carried out with encouraging results.
- This task is 10% completed.

REPORTABLE OUTCOMES:

The PI has produced five first-author abstracts/publications since the original grant submission. These include two conference presentations, one manuscript in-submission, and two published peer-reviewed manuscripts. Unfortunately, only the last three publications credited the present grant, due to late onset of the actual award period (June 2003).

[1] A.O. Bilska-Wolak, C.E. Floyd Jr., Loren W. Nolte, Joseph Y. Lo, "Application of likelihood ratio to classification of mammographic masses; performance comparison to case-based reasoning." Med. Phys. 30 (5), May 2003, pp. 949-958.
[2] A.O. Bilska-Wolak, C.E. Floyd, Jr., Joseph Y. Lo, " Prediction of breast biopsy outcome using a likelihood ratio classifier and biopsy cases from two medical centers." SPIE Medical Imaging, Vol. 5032, p. 1386-1391. 2003.
[3] A.O. Bilska-Wolak, C.E. Floyd Jr., Joseph Y. Lo, "Improved sensitivity for breast cancer classification using a case-based likelihood ratio." MIPS 2003, Durham NC, September 2003.
[4] A.O. Bilska-Wolak, C.E. Floyd Jr, "Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer," Phys. Med. Biol. 49, September 2004,pp. 4219-4237.
[5] A.O. Bilska-Wolak, C.E. Floyd Jr., Joseph Y. Lo, "Computer Aid for Decision to Biopsy Breast Lesions: Pre-clinical Performance Evaluation," (in submission).

- Acquisition of a Ph.D. degree by the principal investigator.
- A large database of BI-RADS™ descriptions for mammography cases from several medical institutions.

CONCLUSIONS:

A first generation likelihood ratio classifier was developed for breast biopsy classification verifying the initial hypothesis.

The performance of the classifier was comparable to or better than other classifiers previously developed for breast biopsy classification. An independent validation test on 151 cases showed that the classifier was able to identify 26% of benign mass lesions that should not be sent to biopsy, while still correctly diagnosing 100% of malignancies. The performance of the classifier was robust even with some missing case data, allowing full utilization of all the information present in the databases.

By decreasing the number of benign cases sent to biopsy, the classifier could be a valuable tool for physicians and ultimately beneficial to hospitals and patients.

REFERENCES:

[1] H. L. VanTrees, Detection, Estimation, and Modulation Theory (Part I), (John Wiley & Sons, New York, 1968).

[2] R. N. McDonough, and A. D. Whalen, Detection of Signals in Noise, (Academic Press, San Diego, 1995).

[3] J. P. Egan, Signal detection theory and ROC analysis, (Academic Press, New York, 1975).

[4] M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, "Ideal observer approximation using Bayesian classification neural networks," IEEE Trans. Med. Imaging **20**, 886-899 (2001).

[5] H. H. Barrett, C. K. Abbey, and E. Clarkson, "Some unlikely properties of the likelihood ratio and its logarithm," in SPIE Med. Imaging: Image Perception, (1998), p.65-77.

[6] A. O. Bilska-Wolak, C. E. Floyd Jr, L. W. Nolte, and J. Y. Lo, "Application of likelihood ratio to classification of mammographic masses; performance comparison to case-based reasoning.," Medical Physics **30**, 949-958 (2003).

[7] A. O. Bilska-Wolak, and C. E. Floyd Jr, "Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer," Phys. Med. Biol. **49**, 4219-4237 (2004).

[8] D. W. Scott, Multivariate density estimation: theory, practice, and visualization, (John Wiley & Sons, New York, 1992).

[9] D. W. Scott, "On optimal and data-based histograms," Biometrica **66**, 605-10 (1979).

[10] M. P. Wand, "Data-based choice of histogram bin width, Australian Graduate School of Management Working Paper Series No. 95-011," University of New South Wales, 1995.

[11] A. O. Bilska-Wolak, C. E. Floyd Jr, and J. Y. Lo, "Prediction of breast biopsy outcome using a likelihood ratio classifier and biopsy cases from two medical centers," in SPIE Medical Imaging, San Diego, CA (SPIE, San Diego, CA, 2003).

[12] A. O. Bilska-Wolak, C. E. Floyd Jr, and J. Y. Lo, "Computer Aid for Decision to Biopsy Breast Lesions: Pre-clinical Performance Evaluation (in submission),".

Appendix A

A.O. Bilska-Wolak, C.E. Floyd, Jr., Joseph Y. Lo, "Prediction of breast biopsy outcome using a likelihood ratio classifier and biopsy cases from two medical centers." SPIE Medical Imaging, Vol. 5032, p. 1386-1391. 2003.

Prediction of breast biopsy outcome using a likelihood ratio classifier and biopsy cases from two medical centers.

Anna O. Bilska-Wolak[a], Carey E. Floyd, Jr.[a,b], Joseph Y. Lo[a,b]
[a]Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA
[b]Department of Radiology, Duke University Medical Center, Durham, NC 27710, USA

ABSTRACT

Potential malignancy of a mammographic lesion can be assessed using the mathematically optimal likelihood ratio (LR) from signal detection theory. We developed a LR classifier for prediction of breast biopsy outcome of mammographic masses from BI-RADS findings. We used cases from Duke University Medical Center (645 total, 232 malignant) and University of Pennsylvania (496, 296). The LR was trained and tested alternatively on both subsets. Leave-one-out sampling was used when training and testing was performed on the same data set. When tested on the Duke set, the LR achieved a Received Operating Characteristic (ROC) area of 0.91± 0.01, regardless of whether Duke or Pennsylvania set was used for training. The LR achieved a ROC area of 0.85± 0.02 for the Pennsylvania set, again regardless of which set was used for training. When using actual case data for training, the LR's procedure is equivalent to case-based reasoning, and can explain the classifier's decisions in terms of similarity to other cases. These preliminary results suggest that the LR is a robust classifier for prediction of biopsy outcome using biopsy cases from different medical centers.

Keywords: computer-aided diagnosis, likelihood ratio, mammography, breast cancer, masses, biopsy

1. INTRODUCTION

The most widely available, reliable and cost-effective method for the early detection of breast cancer is screening mammography.[1] While screening mammography has become a very *sensitive* method for detecting breast abnormalities, the diagnostic process of evaluating a suspicious abnormality is not very *specific*. As many as 70-85% of breast biopsies are performed on benign lesions.[2, 3] The drawbacks of such potentially unnecessary biopsies include stress to patient[4, 5] and increased cost to patient and clinic, which raise the overall cost of screening.[6] Biopsy can also introduce possible distortion on future mammograms, which could make future diagnosis even more difficult.[7] The need to reduce the rate of biopsies performed on benign lesions is well recognized.

The classification of suspicious mammographic lesions by radiologists can be augmented with a second opinion offered by a computer tool.[8, 9] Such computer aids are an inexpensive and a completely non-invasive means of improving the diagnosis. The improvement in diagnosis for the *classification* task in mammography would result in the decrease in the number of biopsies performed on benign lesions. To facilitate this improvement in diagnosis, we have developed a computer classifier based on the likelihood ratio (LRb) for breast biopsy outcome prediction. The LRb's goal is to maintain high sensitivity to malignant lesions, while reducing the number of biopsies performed on benign lesions. The LRb was trained using actual retrospective biopsy cases. The cases were described using the Breast Imaging and Reporting Data System (BI-RADS^{TM}) lexicon,[10] which is the standard for mammography reporting.

Since the features used for classification in this study are part of the standard for mammography reporting, it is conceivable that the one classifier could be used at multiple medical centers. Given a computer-aid trained on data from one medical center, it is therefore imperative to determine whether the aid could be used in another medical center. This study aims to initially answer this question by examining the performance of a likelihood ratio classifier on data from two medical centers.

2. METHODOLOGY

2.1 Data

The database used for this study consisted of non-palpable, proven biopsy cases from two medical institutions. A data set of mammographic mass cases from the Duke University Medical Center consisted of 645 biopsy cases. Of these, 232 were malignant and 413 proved benign at biopsy. Another data set from the University of Pennsylvania Medical center consisted of 496 mammographic mass cases. In this subset, there were 296 malignant cases, and 200 benign.

For this study, each case in the database was represented by two features. One of the features was mass margin, described in accordance with the BI-RADS™ lexicon.[10] The description of mass margin was assigned by experienced mammographers at the time of the decision to biopsy. Mass margin can range from "well circumscribed" to "spiculated." The other feature used was patient age, which is a well known risk-factor for breast cancer. [11]

2.2 Description of classifier

The likelihood ratio (λ) is the optimal detector to determine the presence or absence of signal in noise.[12-14] The signal in this application is the potential malignancy. The decision of whether the malignancy is present or not is optimized by thresholding the likelihood ratio, given by

$$\lambda(X) = \frac{p(X \mid H_1)}{p(X \mid H_0)}$$

where $p(X \mid H_1)$ is the probability density function (PDF) of the features under the "signal present" H_1 hypothesis, and $p(X \mid H_0)$ is the PDF under the "no signal" H_0 hypothesis. The likelihood ratio is optimal under the assumption that the PDFs accurately reflect the true densities. For our likelihood ratio-based (LRb) classifier, the PDFs of the features were estimated using a histogram approach. Since the possible number of findings for BI-RADS™ mass margin is only five, five bins were used to bin mass margin. This binning approach should work well for mass margin, since it might be unnatural to average categorical features. A different approach was used to bin age due to the large possible range of age values. For the age distribution, Scott's rule[15, 16] was first used to determine the optimal bin width. The range of ages within two standard deviations of the mean was divided by the optimal width to determine the bins. The ages falling outside two standard deviations of the mean were included in extreme left or right bins.

2.3 Evaluation

The performance of the classifier was evaluated using Receiver Operating Characteristic (ROC) analysis. We are most interested in the performance at high sensitivities, because missing a malignancy is a graver risk to the patient. Therefore, the main performance metrics included: (1) area under the curve (AUC), which represents average sensitivity over all specificities, (2) partial area[17] index (PAI) above the sensitivity of 90%, which represents the performance of classifier at high sensitivities, and (3) the specificity at 98% sensitivity. The last metric helps indicate how many benign cases could be potentially spared from biopsy while correctly diagnosing 98% of malignancies. This means that 2% of the malignancies could have potentially delayed biopsy and treatment. Since the cases in our database are diagnostic mammography cases, and the physician is already aware of the suspicious lesion, it is conceivable that the performance of the classifier with the physician could be 100% sensitivity.

In order to evaluate the performance of the classifier on the data from the two medical centers, we used four evaluation approaches. (A) The LRb was trained and tested on Duke data, (B) LRb was trained on Pennsylvania data, tested on Duke data, (C) LRb was trained and tested on Pennsylvania data, and (D) LRb was trained on Duke data, tested on Pennsylvania data.

Since several methods exist for training and testing on the same data, approach A was further separated into approaches A_l and A_n. In approach A_l, a leave-one-out sampling approach was applied to evaluate the performance of the Duke-trained classifier on the Duke data set. Each case from the Duke set was used as a test case, while the remainder of the Duke cases was used as the training set. This was repeated over all cases, until all cases had been used as a test case. The ROC curve was produced by thresholding the likelihood ratio values of the test cases. In approach A_n, the classifier was trained and tested on all cases. This produced a consistency curve for the Duke set.
A similar procedure was followed for approach C, for training and testing on the Pennsylvania data set. The leave-one-

out approach approach is referred to as C_l, and the consistency approach is referred to as C_n.

For approach D (cross-training across different institutions) all of the Duke data was used for training the classifier, and then all of the Pennsylvania data was used for testing. In a similar fashion for approach B, all of the Pennsylvania data was used for training the classifier, and then all of the Duke data was used for testing performance. In each situation, all of the training dataset was used to establish the threshold which would spare 98% of the benign cases from the training dataset. For approach D, the 98% sensitivity threshold from approach A_n was used. This means that the threshold was established using all of the Duke cases (the consistency curve). This threshold was then applied to the ROC curve of the test set. From this ROC curve, the new resulting sensitivity and specificity on the test were acquired. These results are presented in Table 2. Similarly for approach B, the 98% sensitivity threshold from approach C_n was used to establish the new sensitivity and specificity on the test set.

2.4 Relationship to case-based reasoning

Case-based reasoning (CBR) is an artificial intelligence approach for solving problems based on the premise that similar problems have similar solutions.[18] The CBR is based on the human learning process - learning by experience. A CBR system can illustrate which cases from a database were found similar to a new case, thus presenting compelling justification for the system's diagnostic outcome.[19] This property is important for potential clinical use, since physicians and patients are uncomfortable receiving a diagnostic decision without an explanation behind it. In our studies, we have noticed that the CBR is a likelihood ratio based algorithm. The CBR in effect uses similarity metrics to approximate and smooth the feature distributions. From this relationship, we have been able to realize that many likelihood ratio based classifiers can also explain their reasoning in terms of similar cases. This can make the LRb a more attractive classifier for a real clinical situation, since the classifier can explain its diagnostic decision to a physician. The LRb can return to the user cases that were found similar to the case in question, their mammograms, BI-RADS™ findings, and other relevant information, along with the diagnostic decision.

3. RESULTS

3.1 Testing the trained classifier on the Duke dataset (Approaches A-B)

When the Duke dataset was used for training and testing the classifier in a leave-one-out fashion, the AUC was 0.91 ± 0.01 (Table 1). When the Penn dataset was used for training the classifier, the performance of the classifier tested on the Duke set was again 0.91 ± 0.01. While no difference was observed in the AUC, there was a small difference in the PAI, above the 90% sensitivity level. The PAI for testing and training on Duke was 0.59 ± 0.05, while it was only 0.53 ± 0.07 for training on Penn and testing on Duke. The PAI was lower for testing and training on data across different medical centers, while the AUC remained the same. The difference in PAI, was not statistically significant (p=0.57).

Table 1: Performance of the classifier on Duke and Pennsylvania datasets.

Approach	Training Dataset	Testing Dataset	AUC	PAI (above 90% sensitivity)	# Benign Cases Potentially Spared at 98% Sensitivity
A_l	Duke	Duke (leave-one-out)	0.91 ± 0.01	0.59 ± 0.05	209 (51%)
A_n	Duke	Duke (consistency)	0.94 ± 0.01	0.71 ± 0.03	264 (64%)
B	Pennsylvania	Duke	0.91 ± 0.01	0.53 ± 0.07	see Table 2
C_l	Pennsylvania	Pennsylvania (leave-one-out)	0.85 ± 0.02	0.30 ± 0.07	28 (14%)
C_n	Penn	Penn (Consistency)	0.90 ± 0.01	0.52 ± 0.06	78 (39%)
D	Duke	Pennsylvania	0.85 ± 0.02	0.29 ± 0.07	see Table 2

To examine the potential clinical benefits, we examined the number of benign cases that could be spared from biopsy. Using the leave-one-out classifier trained on the Duke dataset (at 98% sensitivity) could potentially spare from biopsy 51% of the benign cases from the Duke dataset. Training on the Pennsylvania dataset and using the 98% threshold established on the Pennsylvania set would also result in 98% sensitivity on the Duke set, and potentially obviate 46% benign cases from Duke University. It would be slightly more beneficial to train and test on data from the same medical institution, since the LRb could potentially obviate 51% versus 46% of the benign biopsies at the 98% sensitivity level.

Table 2: Performance of the cross-trained classifier using 98% sensitivity threshold established on the training data.

Approach	Training Dataset	Testing Dataset	Actual Resulting Sensitivity	# Benign Cases Potentially Spared from Biopsy
B	Pennsylvania	Duke	98%	189 (46%)
D	Duke	Pennsylvania	92%	104 (52%)

3.2 Testing the trained classifier on the Pennsylvania dataset (Approaches C-D)

When the Pennsylvania dataset was used for training and testing the classifier, the leave-one-out AUC was 0.85 ± 0.02. When the classifier was trained on the Duke dataset, and tested on the Pennsylvania dataset, the AUC was also 0.85 ± 0.02. Therefore, the performance of the classifier tested on the Pennsylvania data was the same in terms of AUC, regardless of which dataset was used for training. While no difference was observed in the AUC, there was a small difference in the PAI. The PAI for testing and training on the Pennsylvania data was 0.30 ± 0.07, and 0.29 ± 0.07 for training on the Duke data. This difference was also not statistically significant (p=0.99).

After training the classifier in a leave-one-out fashion on the Pennsylvania data, the LRb could potentially spare from biopsy 14% of the Pennsylvania benign lesions. Training on the Duke dataset and using the 98% threshold established on the Duke set would result in 92% sensitivity on the Pennsylvania data set. This sensitivity was also the next highest sensitivity point after 100% on the Pennsylvania curve. Using the classifier at this sensitivity would potentially obviate 46% benign cases from the Pennsylvania data set. These results suggest that while it is possible to train and test on data from different medical centers, it may be more beneficial to train and test on data from the same medical center.

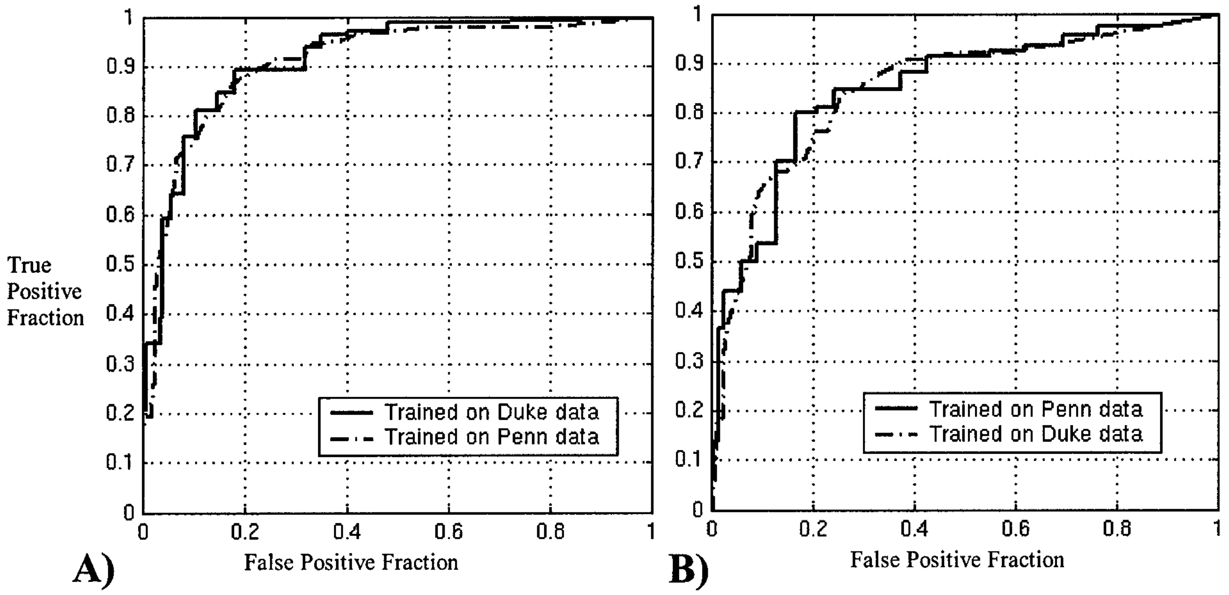

Figure 1: Performance of the classifier tested on on A) data from Duke University, and B) data from University of Pennsylvania

4. DISCUSSION

The high performance of the LRb classifier using only two of the findings suggests that a lot of information is present in the age distribution, and in the mass margin finding as assigned by mammographers.
It is difficult to ascertain why the performance of the classifier was lower on testing on the Pennsylvania data than on the Duke data. Given that the performance remains comparable regardless of which dataset is used for *training*, it is likely these databases do not contain more information for improvement of results.

It is difficult to ascertain why training on the Duke dataset and establishing the 98% sensitivity threshold did not produce equally high sensitivity in the testing on the Pennsylvania set. The threshold established on the Duke set resulted in a sensitivity drop to 92% on the Pennsylvania dataset. This sensitivity was also the next highest sensitivity point after 100% on the Pennsylvania curve. Since these results did not involve resampling of the inputs, more research is needed to examine why this drop in sensitivity was observed. However, it is encouraging that training the classifier on the Pennsylvania database did not result in a sensitivity drop. The 98% threshold established on the Pennsylvania set resulted in equally high sensitivity of 98% on the Duke set. The classifier would still be able to maintain 98% sensitivity on the Duke data at the training threshold, while potentially sparing from biopsy 46% of the benign cases.

While the performance of the classifier decreased at high sensitivities when training on a dataset from another medical center, the overall performance (AUC) remained the same regardless of which dataset was used for training. These preliminary results suggest that it may be possible to use data from multiple medical centers to create a global classifier for breast cancer prediction.

The LRb remains an attractive classifier, since it can result in optimum performance and can explain reasoning behind its decision in terms of similar cases. The LRb can return to the user cases that were found similar to the case in question, their mammograms, BI-RADS™ findings, and other relevant information, along with the diagnostic decision. Furthermore, the performance of the classifier remained comparable regardless of which database was used for training. These preliminary results suggest that the LRb is a robust classifier for prediction of breast biopsy outcome using biopsy cases from different medical centers.

ACKNOWLEDGMENTS

This work was supported in part by USPHS grant number R21-CA81309, awarded by the National Cancer Institute; USAMRMC grant number DAMD17-99-1-9174 awarded by the U.S. Army, and Susan G. Komen Breast Cancer Foundation grant number BCTR2000730A.

REFERENCES

[1] S. A. Feig, "Role and evaluation of mammography and other imaging methods for breast cancer detection, diagnosis, and staging," Sem. Nucl. Med. **29**, 3-15 (1999).
[2] D. B. Kopans, "The positive predictive value of mammography," AJR **158**, 521-526 (1992).
[3] J. E. Meyer, T. J. Eberlein, P. C. Stomper, and R. R. Sonnefeld, "Biopsy of occult breast lesions: Analysis of 1261 abnormalities," J. Am. Med. Assoc. **263**, 2341-43 (1990).
[4] M. A. Helvie, D. M. Ikeda, and D. D. Adler, "Localization and needle aspiration of breast lesions: complications in 370 cases," AJR **157**, 711-714 (1991).
[5] J. M. Dixon, and T. G. John, "Morbidity after breast biopsy for benign disease in a screened population," Lancet **1**, 128 (1992).
[6] D. Cyrlak, "Induced costs of low-cost screening mammography," Radiology **168**, 661-3 (1988).
[7] M. D. Kaye, M. L. Vicinanza-Adami, and M. L. Sullivan, "Mammographic findings after stereotaxic biopsy of the breast performed with large-core needles," Radiology **192**, 149-151 (1994).

[8] Y. Jiang, R. M. Nishikawa, R. A. Schmidt, A. Y. Toledano, and K. Doi, "Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications," **220**, 787-794 (2001).

[9] K. Doi, H. MacMahon, S. Katsuragawa, R. M. Nishikawa, and Y. Jiang, "Computer-aided diagnosis in radiology: potential and pitfalls," **31**, 97-109 (1999)(review).

[10] BI-RADS, "American College of Radiology Breast Imaging - Reporting and Data System (BI-RADS) 3rd ed.," American College of Radiology, 1998.

[11] B. Hulka, and P. Moorman, "Breast cancer: hormones and other risk factors," Maturitas **38**, 103-113 (2001).

[12] H. L. VanTrees, Detection, Estimation, and Modulation Theory (Part I), (John Wiley & Sons, New York, 1968).

[13] R. N. McDonough, and A. D. Whalen, Detection of Signals in Noise, (Academic Press, San Diego, 1995).

[14] J. P. Egan, Signal detection theory and ROC analysis, (Academic Press, New York, 1975).

[15] D. W. Scott, Multivariate density estimation: theory, practice, and visualization, (John Wiley & Sons, New York, 1992).

[16] D. W. Scott, "On optimal and data-based histograms," Biometrica **66**, 605-10 (1979).

[17] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," Radiology **201**, 745-750 (1996).

[18] D. B. Leake, Case-based reasoning: experiences, lessons, and future directions, (AAAI Press, Menlo Park CA, 1996).

[19] I. Watson, Applying Case-Based Reasoning: Techniques for Enterprise Systems, (Morgan Kaufmann Publishers, San Francisco CA, 1997).

Appendix B

A.O. Bilska-Wolak, C.E. Floyd Jr, "Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer," Phys. Med. Biol. 49, September 2004,pp. 4219-4237.

Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer

Anna O Bilska-Wolak[1] **and Carey E Floyd Jr**[1,2]

[1] Department of Biomedical Engineering, Duke University, 2623 DUMC, Durham, NC 27708, USA

[2] Department of Radiology, Duke University Medical Center, 2623 DUMC, Durham, NC 27710, USA

E-mail: anya.bilska@duke.edu

Received 8 April 2004, in final form 21 June 2004
Published 27 August 2004
Online at stacks.iop.org/PMB/49/4219
doi:10.1088/0031-9155/49/18/003

Abstract

While mammography is a highly sensitive method for detecting breast tumours, its ability to differentiate between malignant and benign lesions is low, which may result in as many as 70% of unnecessary biopsies. The purpose of this study was to develop a highly specific computer-aided diagnosis algorithm to improve classification of mammographic masses. A classifier based on the likelihood ratio was developed to accommodate cases with missing data. Data for development included 671 biopsy cases (245 malignant), with biopsy-proved outcome. Sixteen features based on the BI-RADS$^{\mathrm{TM}}$ lexicon and patient history had been recorded for the cases, with 1.3 ± 1.1 missing feature values per case. Classifier evaluation methods included receiver operating characteristic and leave-one-out bootstrap sampling. The classifier achieved 32% specificity at 100% sensitivity on the 671 cases with 16 features that had missing values. Utilizing just the seven features present for all cases resulted in decreased performance at 100% sensitivity with average 19% specificity. No cases and no feature data were omitted during classifier development, showing that it is more beneficial to utilize cases with missing values than to discard incomplete cases that cannot be handled by many algorithms. Classification of mammographic masses was commendable at high sensitivity levels, indicating that benign cases could be potentially spared from biopsy.

1. Introduction

Mammography is the most readily available modality for the early detection of breast cancer. Clinical success of mammography is due to its high sensitivity to malignant lesions combined with cost-effectiveness. Several comprehensive studies have demonstrated the efficacy of screening mammography on the general population of asymptomatic women (Tabar *et al* 2001, Morimoto *et al* 2000, Miller *et al* 1992). While screening mammography concentrates on the detection of abnormalities in the general female population, diagnostic mammography concentrates on classifying encountered abnormalities and recommending an appropriate course of medical action.

The high sensitivity of screening mammography is compromised by its low specificity to benign lesions which often appear mammographically similar to malignant lesions. This results in approximately 70% of biopsies (Kopans 1992, Meyer *et al* 1990) performed on benign lesions. The negative effects of potentially unnecessary biopsies include pain, anxiety, altered cosmetic appearance and monetary cost (Helvie *et al* 1991, Dixon and John 1992). Biopsy can also introduce distortion on future mammograms, which could complicate diagnosis in the future. Regular screening for women starting at the age of 40 is recommended by the American College of Radiology (ACR). Approximately 64 million women in the Unites States are above the age of 40 (US Census Bureau 2001). Therefore, potentially unnecessary biopsies might be performed on millions of women (Kaye *et al* 1994). It would be beneficial both to individuals and society as a whole to reduce the number of biopsies performed on benign lesions.

Radiologists' diagnostic performance in mammography can be enhanced by computers providing a rapid and accurate second opinion using appropriate algorithms (Jiang *et al* 2001, Doi *et al* 1999). Such algorithms are a potentially noninvasive, low-cost solution to improving the diagnosis. Currently, mammography is still under-utilized by the female population (Michaelson *et al* 2002) and the number of dedicated-breast imaging radiologists is often unsatisfactory. If the utilization of mammography improves, screening resources are likely to become more inadequate in the future. This increased need for improved mammographic screening can be partially alleviated by computer-aided diagnosis (CAD).

CAD refers to the utilization of a computer tool by a physician to help in medical diagnosis. CAD for breast cancer generally pertains to the use of computers by dedicated breast-imaging radiologists to assist in detection and/or classification of suspicious lesions present on mammograms. These lesions include masses, calcifications and architectural distortions. Various CAD algorithms have been applied to classification (Hadjiiski *et al* 2001, Lo *et al* 1999, Veldkamp *et al* 2000) and detection (Gavrielides *et al* 2000, Chang *et al* 2001) of suspicious lesions. CAD for detection refers to identifying suspicious lesions on mammograms, and classification refers to deciding whether the suspicious lesion is malignant or benign.

Physicians describe mammographic lesions using the standard lexicon from the ACR Breast Imaging and Data Reporting System (BI-RADS$^{\text{TM}}$) (BI-RADS 1998). Information from the BI-RADS$^{\text{TM}}$ lexicon has been used previously to improve classification of breast cancer (Floyd *et al* 2000) by employing ordered rankings that were assigned to the categorical descriptions of BI-RADS$^{\text{TM}}$. However, the best way of encoding categorical features into mathematical representation for use by a classifier is not obvious and an inadequate implementation can have an adverse effect on performance. This problem will be addressed by this study.

Another problem that is frequently encountered during the development of a CAD tool for medical purposes is the lack of complete data. This lack of complete data may present itself an inadequately small number of cases for development, or a proportion of cases that have missing values and are thus unsuitable for most classification algorithms. In mammography, for example, an algorithm may be developed that employs knowledge of the patient hormone-therapy history that is not part of the BI-RADS$^{\text{TM}}$ lexicon. Hormone-therapy history might not have been collected for certain cases in the available database. If one requires the algorithm to use hormone-therapy history, the lack of this information for some cases can reduce the total number of cases available for the development of the computer-aid. This lack of values, leading to a reduced number of cases, can prevent one from utilizing all information available in the database.

In this investigation, we present an improvement to a previously developed classifier for breast cancer prediction. While the initial classifier was developed using cases with complete data, the present classifier was developed to accommodate cases with missing values. This allowed us to incorporate more features from patient medical history and also more cases. The utilization of all information in the database produced a more sensitive and specific classifier. We demonstrate the performance of the classifier on a very large data set, and exemplify its robust performance with the missing values. We achieve excellent performance for breast biopsy prediction at the 100% sensitivity level.

2. Methods

2.1. Description of database

The BI-RADS$^{\mathrm{TM}}$ data were collected as part of standard clinical practice at Duke University Medical Center, in accordance with institutional safety review boards. The data were compiled within several discontinuous time periods between 1991 and 2000, but consecutively within each time period. The original data consisted of 1530 non-palpable biopsy-proved cases. Of these, 61 cases were removed because it was not possible to confirm that they were non-palpable, leaving 1469 cases (512 malignant). In this study, only mass cases were examined. Mass cases were defined as cases that had a mass and any other findings, such as associated calcifications or architectural distortions. There were 671 mass cases in this database, and 245 proved to be malignant at biopsy. Biopsy outcome for each case was obtained from the histopathological analysis.

The cases in the database were represented by features based on the BI-RADS$^{\mathrm{TM}}$ lexicon and patient history findings. These 16 features included mass margin, mass shape, mass density, mass size, calcification morphology, calcification distribution, calcification number, associated findings, architectural distortion, special findings and patient age. The previously unused features were: change from prior mammogram, hormone-therapy history, breast side, quadrant location and menopausal status. The possible descriptors (findings) for each feature are listed in table 1. (Nomenclature: note that mass margin is referred to as a 'feature', while microlobulated is referred to as 'finding' or 'feature value'.)

Many of the cases present in the database had some feature values unrecorded. Please note the difference between unrecorded features, and the absence of findings for a specific case. For example, a suspicious mass might be located on a mammogram without associated calcifications. Therefore, the calcification findings will say 'no calcifications'. However, if a suspicious mass does have calcifications, and for unknown reasons the calcification morphology was not recorded, then the value for that feature is missing. It is then specified as 'no information' finding for calcification morphology. For the 671 cases, 1.3 ± 1.1 feature values were missing on average per case (table 2). Each feature had a different rate of scarcity in the data, and only seven features were present for all 671 cases. The seven features recorded for all the 671 cases included mass margin, calcification number, special findings, associated findings, patient age, breast side and architectural distortion. All 16 feature values were recorded for only 225 cases (81 malignant) in the database. Unrecorded feature values will be henceforth referred to as 'missing'.

Table 1. Mammographic features and their findings as used by the classifier[a].

	Feature	Finding		Feature	Finding
(1)	Patient age	Years	(9)	Calcification number	No calcs
					<5
(2)	Mass margin	No mass			5 to 10
		Well circumscribed			>10
		Microlobulated			No info
		Obscured	(10)	Associated findings	None
		Ill-defined			Skin lesion
		Spiculated			Haematoma
		No info			Post-surgical scar
(3)	Mass density	No mass			Trabecular thickening
		Fat-containing			Skin thickening
		Low density			Skin retraction
		Isodense			Nipple retraction
		High density			Axillary adenopathy
		No info			Architectural distortion
(4)	Mass shape	No mass			No info
		Round	(11)	Special cases	None
		Oval			Intrammamary lymph node
		Lobular			Assymetric breast tissue
		Irregular			Focal assymetric density
		No info			Tubular density or solitary dilated duct
(5)	Mass size	mm			
					No info
(6)	Menopause	Pre-menopausal	(12)	Quad	Posterior
		Post-menopausal			Central
		No info			LIQ
(7)	Calcification distribution	No calcs			LOQ
		Diffuse			UIQ
		Regional			UOQ
		Segmental			Axillary tail
		Linear			Subareolar
		Clustered			No info
		No info	(13)	Change from prior	No change
(8)	Calcification morphology	No calcs			New lesion
		Milk of calcium-like			Qualitative change
		Eggshell or rim			Quantitative change
		Skin			No info
		Vascular	(14)	Architectural distortion as main finding	None
		Spherical or lucent-centred			Present
		Suture			No info
		Coarse ('popcorn')	(15)	Hormone use	No hormone use
		Large rod-like			Hormones (such as BCP oestrogen, progesterone)
		Round			
		Dystrophic			No info
		Punctate	(16)	Breast side	Left
		Indistinct			Right
		Pleomorphic			No info
		Fine branching			
		No info			

[a] The percentages of missing values are as follows: (1) 0%, (2) 0%, (3) 2%, (4) 1%, (5) 4%, (6) 40%, (7) 2%, (8) 1%, (9) 0%, (10) 0%, (11) 0%, (12) 3%, (13) 14%, (14) 0%, (15) 40% and (16) 0%.

2.2. *The likelihood ratio based algorithm*

Given features describing the presence of a suspicious lesion on a mammogram, one has to decide whether or not a malignancy—a signal—is present. The optimal detector to

Table 2. Number of cases in database with respect to number of features and missing values. Please note that B, C, and A_7 are actually subsets of A, and that A = B + C.

Set ID	Number of cases	Number of features	Number of missing feature values per case
A	671	16	1.3 ± 1.1
A_7	671	7	0
B	225	16	0
C	446	16	1+

determine the presence or absence of a signal in noise is the likelihood ratio (VanTrees 1968, McDonough and Whalen 1995, Egan 1975, Kupinski *et al* 2001, Barrett *et al* 1998). The likelihood ratio λ is optimal under the assumption that full knowledge of the statistical properties of the data is known. The likelihood ratio is also referred to as the ideal observer.

The hypothesis that the signal is present is the H_1 hypothesis, and the 'no signal' or null hypothesis is referred to as H_0. The decision whether the signal is present or absent is optimized by thresholding the likelihood ratio $\lambda(V)$,

$$\lambda(V) = \frac{p(V|H_1)}{p(V|H_0)}$$

where the $p(V|H_1)$ represents the probability density function (PDF) of the available data V under the 'signal present' hypothesis, and $p(V|H_0)$ represents the PDF of the available data under the 'no signal' hypothesis. In this study, the signal to be detected is the malignancy of the breast lesion. Therefore, we can use available descriptions for *malignant* biopsy cases to represent the H_1 distribution, and the descriptions for *benign* biopsy cases to represent the H_0 distribution. The next challenge concerns the logistics of the representation of the H_0 and H_1 feature distributions for the biopsy cases. The biopsy cases present in our database were described with BI-RADS$^{\text{TM}}$ lexicon, which is a collection of categorical descriptives for features (table 1). In past studies, the categorical descriptives for each feature were arranged in order of increasing risk of malignancy, and a ranking scale was assigned (Lo *et al* 1999, Floyd *et al* 2000, Bilska-Wolak and Floyd 2002a). This scale had been established in consultation with physicians, and has also been used with artificial neural networks (Lo *et al* 1999), linear discriminant analysis (Markey *et al* 2002), case based reasoning (Floyd *et al* 2000, Bilska-Wolak and Floyd 2002a), and constraint-satisfaction neural networks (Tourassi *et al* 2001). While the ranking approach is practical and has resulted in commendable performance, it is possible that better performance can be achieved without the use of a classifier that depends on a numerical or ranking scale. This is true because while one finding might be considered more malignant than another, real-life performance and database content might not be reflected in the ranking assignments. For example, physicians might consider *dystrophic* calcifications more at risk to be malignant than *round*, and thus designate a rank of 10 to *dystrophic* calcifications, and 9 to *round* calcifications. However, physicians might assign *round* more often to malignant calcifications than *dystrophic*, thus behaving in opposition to their scale. Any classifier that utilizes averaging or interpolation of findings, for example, will be dependent on this scale/ordering of the categorical features, and may be affected unfavourably.

For our likelihood ratio based algorithm, we have chosen to represent the categorical features as discrete histogram distributions. Such nonparametric models can be very effective, and also reduce the risk of misinterpreting the data (Scott 1992). Conversely, parametric models often interpolate the data, yet it might be unnatural to average categorical feature findings. Given the small number of findings for the BI-RADS$^{\text{TM}}$ features, we can represent the densities as histograms with fixed bins. This will allow the findings of a feature to remain separate and unaffected by ordering. For example, since mass shape has only four possible descriptions, four bins will be designated for each finding. We have also introduced a fifth bin, to represent the lack of information, or 'no info'. Therefore, the mass shape histogram has five possible bins. A different strategy was chosen for the continuous findings, such as age and mass size. Creating a histogram for these is slightly more complex due to the large number of possible values. We used Scott's rule (Scott 1992, 1979, Wand 1995) to determine the optimal histogram bin width h. Let σ represent the standard deviation and n the number of available observations. The optimal bin width h is then,

$$h = 3.5 \times \sigma \times n^{-1/3}.$$

The interval within two standard deviations of the mean was subdivided by the bin width. Observations falling outside the two standard deviations were included in extreme right and left bins. For the mass size distribution, this resulted in nine bins. An extra bin indicating 'no information' was also added, resulting in ten bins for the mass size distribution. Essentially, in this set-up the encoding of categorical (BI-RADS$^{\text{TM}}$) findings matters a little. The only fact of importance is that each finding remains a separate category. We eventually encode the findings as numbers in the algorithm, because it is easier for a computer program to use them, but the values we choose can be any number. This can be accomplished due to the way we have defined our distributions and how the likelihood ratio is computed.

While it would be ideal to compute the f-variate distribution histogram of all the features for the likelihood ratio *based* (LRb), populating the 16-dimensional feature space presents a problem: an extremely large number of cases would be needed to provide an adequate representation of the 16-dimensional space. One solution to this problem consists of using an assumption of feature independence and creating separate likelihood ratio classifiers for each feature. By merging the outputs from the individual feature classifiers, the final classifier can then be created. The final merged classifier is not a true likelihood ratio classifier, but strictly speaking, is a likelihood ratio *based* classifier. Our merging method consisted of summing over all individual likelihood ratios. We have found this to work better than multiplication or any other merging method. This result may suggest that summing reduces the influence of extreme likelihood ratio values better than multiplication. These very small or very large values may be the result of a small number of cases, potentially suggesting less accurate local PDF estimation and thus less accurate likelihood ratio estimation. (Note that averaging is equivalent to summing, since it is a monotonically increasing transformation of the likelihood ratio.) The merging of individual features implies feature independence, but the assumption appears to work well for our purpose. For example, the linear correlation coefficient between features for all cases is very low (less than $|0.2|$) for majority of feature pairs, implying low (linear) interdependence of features. Furthermore, each feature distribution is estimated using all available training cases, implying a more robust PDF for each feature. The benefit of the summing method has been also shown in other empirical studies. For example, Kittler *et al* (1998) found that the classifier combination rule developed under most restrictive assumptions—the sum rule—outperformed other classifier combination schemes. Zheng *et al* (2001) found that averaging independent observers (neural networks) improved performance for breast mass detection. Swensson *et al* (2000) found gains in performance when averaging radiologists' ratings of abnormality. In effect, the summing method suggests that the more high-malignancy-risk features describe a specific case, the more likely that the case will be predicted as malignant by the LRb.

2.3. LRb versus other classification algorithms for biopsy prediction

Many classifiers for breast biopsy classification based on BI-RADS$^{\mathrm{TM}}$ are likely to have poor performance because of a dependence on the number scale assigned to categorical features. As mentioned earlier, the number scale might not be accurately reflected in the rankings. A classifier often employed in CAD is Fischer's linear discriminant. Fischer's linear discriminant is actually a likelihood ratio classifier with feature distributions estimated using the multivariate normal assumption with equal covariance matrices. Since the multivariate normal assumption might not accurately represent the data, and is dependent on the number scale assigned to features, it may not be appropriate to use this assumption/classifier. Using the multivariate assumption may prove particularly difficult when encoding categorical features. It is also difficult in practice to encode 'no information' into mathematical representation for multivariate normal assumptions. Other classifiers such as neural networks may also be non-optimal for classifying categorical features since such classifiers also depend on the number scale/rankings assigned to feature findings.

Most classification algorithms require the presence of all feature values for the functioning of the algorithm. Therefore, the lack of certain feature values means that parts of the data have to be excluded from algorithm development. Lack of data may be a mere error or omission during the collection process. However, feature values may also be absent due to the physicians' uncertainty, indicating a pattern of assignment. In this situation, lack of a feature value could actually provide useful information. Furthermore, decreasing the number of cases available for algorithm development means that fewer cases are available to estimate the probability density functions of features. Fewer cases could result in less accurate estimation of the probability density functions of the features, which could result in poorer performance. Both of these problems can be circumvented by utilizing a classifier, such as the LRb, that can utilize cases with missing values.

Other methods for dealing with missing data are available. An example method of dealing with missing data is to simply substitute the missing value by the mean of all known values of the variable (Beale and Little 1975, Little and Rubin 2002). This simple method is an often practically utilized method of dealing with missing data. This method is compared with 'missing' information category approach for one feature in the results section.

The LRb has other attractive features. As we have defined it in this manuscript, the LRb has less potential for over-training than many complex classifiers. There is no danger for over-training in potential feature selection since all the available features are used. There is also little potential for over-training in defining the distributions. The distributions are defined using discrete histograms with fixed bins. For the majority of the features (which are categorical), the number of bins is dependent solely on the number of possible findings for that specific feature. For the continuous features, the bins are established using Scott's rule and are also kept fixed. This method also does not use any knowledge of the biopsy outcome. The separation of categorical findings using bins permits the use of cases that are lacking features (since we can have a separate 'no information' bin), thus maximizing the database available for development. The method is also intuitively simple, and employs a 'natural' form of scaling. By 'natural' we mean that feature values that have proportionally a lot malignant cases are weighted more heavily, as they will have a larger value of the likelihood ratio and thus will be considered more likely to be malignant. Additionally, since the actual algorithm is not very computationally intensive, it is possible to train and test the LRb using advanced validation procedures (*vide infra*) to conduct a comprehensive performance evaluation.

2.4. Classifier performance evaluation

Receiver operating characteristic (ROC) analysis was used to evaluate the performance of the classifier. The performance at high sensitivities is of foremost interest in biopsy prediction since missing a malignancy is a greater risk to the patient. The high sensitivity region was not represented well in the parametric binormal model evaluations. Therefore, the non-parametric ROC was computed for all evaluations. Since the ROC is computed non-parametrically, we can obtain simple and clinically useful interpretations of the performance. Points on the non-parametric ROC curve can be used to obtain values of performance at various sensitivities that can be translated into performance in terms of numbers of actual cases.

The main performance measures included: (1) area under the curve (AUC), which represents average sensitivity over all specificities, (2) partial area index (${}_{0.90}$AUC) (Jiang *et al* 1996) above the sensitivity of 90%, which represents the performance of the classifier at high sensitivities, (3) specificity at 95% sensitivity, and (4) specificity at 100% sensitivity. In this application, specificity at 95% sensitivity indicates how many benign cases could be potentially spared from biopsy while correctly diagnosing 95% of malignancies. Ninety-five per cent sensitivity indicates that 5% of the malignancies could have potentially delayed biopsy and treatment. Specificity at 100% sensitivity indicates how many benign cases could be potentially spared from biopsy while correctly diagnosing 100% of malignancies. The sensitivity at a fixed FPF is often preferable to the AUC when evaluating a test for a particular application (Zhou *et al* 2002).

2.5. Classifier evaluation methods

The two essential developmental stages for any classifier are training and testing. In the training stage, the classifier learns from the training data. In the testing stage, the classifier evaluates previously unseen cases, and we examine the resulting performance. Similarly for the LRb algorithm, the training stage is the computation of the likelihood ratios from mammographic features of the training cases. The testing stage consists of extracting the merged likelihood ratio values for given testing cases, and using those likelihood ratio values for computing the ROC curve.

Several popular approaches exist for testing and training a classifier on a limited set of data. These include the round robin, *k*-fold cross-validation, and resubstitution (Kohavi 1995). The bootstrap is another powerful resampling technique (Efron and Tibshirani 1993, Jain *et al* 1987), which is rarely described in medical decision making literature. The bootstrap models the relationship between the true (unknown) distribution, and the sample (our collected data) by the relationship between the sample and a subsample drawn from our collected sample (Hand *et al* 2001). In the classic bootstrap, a population of *n* cases is resampled with replacement to form B bootstrap samples of size n each. Because of replacement, each sample contains some duplication. Each bootstrap sample can be used to compute some statistic (such as the interquartile range). Averaging the statistic value over all bootstrap samples will give the mean and variance for this statistic. For the classification task, where both training and testing data sets are needed, the bootstrap has been applied a little differently. In the simplest application of the bootstrap, each bootstrap sample in turn is used for training the classifier. After each training-run, the classifier is tested on the original sample (Efron and Tibshirani 1993) resulting in B estimates of performance. Averaging over all bootstrap samples results in the final estimate of performance. Similarly to the resubstitution method, this bootstrap is optimistically biased for classification, since there is repetition of cases in the training and testing data sets.

Another, better resampling solution is the leave-one-out bootstrap (Jain *et al* 1987, Efron and Tibshirani 1997). Leave-one-out bootstrap is also sometimes referred to as the ε_0 estimate in the earlier literature, and is a subcomponent of the of 0.632 bootstrap (Efron and Tibshirani 1997). In the leave-one-out bootstrap (as applied to calculating, for example, the classification error rate), the population of n cases is resampled with replacement to form B bootstrap samples of size n each. Each bootstrap sample is used to train the classifier, and then the classifier is tested on the instances that do not occur in the bootstrap sample. The error rate is averaged over all cases and all bootstrap samples to give the final error rate estimate. Let the data X be represented by $\boldsymbol{x}_i = [\boldsymbol{v}_i, y_i]$, for $i = 1, \ldots, n$ cases, where $\boldsymbol{v}_i$ is the vector training data, and y_i is the class membership. Formally, the error rate ε_0 is thus,

$$\hat{\varepsilon}_0 = \frac{\sum_{b=1}^{B} \sum_{Ab} Q[y_i, \eta_{X^{*b}}(\boldsymbol{v}_i)]}{\sum_{b=1}^{B} |A_b|}$$

where A_b is the set of training patterns i that do not appear in the bth bootstrap sample, and $|A_b|$ is the cardinality of A_b. Q is the measure of error between the class membership y_i and the prediction η. The prediction η for the data $\boldsymbol{v}_i$ is derived from training on bootstrap sample X^{*b}. Little information exists in the literature about the theoretical accuracy of applying the leave-one-out bootstrap to ROC analysis. However, leave-one-out bootstrap has been applied in this fashion. In effect, the leave-one-out bootstrap results in the training on approximately 63% of the data, and testing on 37% of the data, repeated B times (Kohavi 1995). Each time a slightly different data set is used for testing and training. The ROC performance is averaged over the B samples. The training and testing data sets for the leave-one-out bootstrap are mutually exclusive, implying some negative performance bias, at least for such point estimates as prediction error (Efron and Tibshirani 1997).

In this study, we used the leave-one-out bootstrap when training and testing were performed on the same subset. When training on one data subset and testing on another subset, separate bootstrap samples were drawn from each subset. Worthy to note is the extreme computational demand when applying the bootstrap. For each bootstrap sample, the classifier has to be retrained anew. This computational overhead is probably the reason why leave-one-out bootstrap has not been applied extensively in ROC or CAD studies previously. In order to introduce the leave-one-out bootstrap for CAD, we have applied all five methods (round robin, resubstitution, cross-validation, bootstrap, leave-one-out bootstrap) to the whole data subset. To simplify the remaining evaluations, only the leave-one-out bootstrap was used on the rest of the data subsets.

The aforementioned methods were used for computing the results. All evaluation, validation and classification methods were custom programmed in MATLAB, and evaluated on a Sun Ultra10 workstation.

3. Results

Section 3.1 examines the performance of the leave-one-out bootstrap on the full dataset and how this evaluation method compares to round robin, cross-validation, resubstitution and bootstrap methods. In section 3.2, we exemplify the results of using and not using the missing information content for one feature. In section 3.3, we investigate the leave-one-out bootstrap performance of the classifier on various subsets of the data that includes or does not include missing values.

Table 3. Results of the classifier trained and tested on 671 cases and all 16 features. Some of the cases have feature values missing.

Evaluation type	AUC ± STD	$_{0.90}$AUC ± STD	Specificity at 95% sensitivity (%)	Specificity at 100% sensitivity (%)
Round robin	0.91 ± 0.01	0.62 ± 0.04	66	26
Cross validation ($k = 10$)	0.91 ± 0.002	0.61 ± 0.007	66	23
Leave-one-out bootstrap	0.91 ± 0.02	0.60 ± 0.06	64	32
Bootstrap	0.92 ± 0.004	0.63 ± 0.01	68	24
Resubstitution	0.93 ± 0.01	0.65 ± 0.04	70	29

3.1. Comparison of leave-one-out bootstrap to other evaluation methods

In order to demonstrate the performance of the leave-one-out bootstrap with respect to the round robin, cross-validation, resubstitution and conventional bootstrap, we have executed all five evaluations on data subset A (defined in table 2). These results for subset A, evaluation on all 671 available cases and 16 features, are located in table 3. Cross-validation was performed 2000 times, with various random splits of the data. For the bootstrap evaluations, 3000 bootstrap samples were used.

The classifier has high AUC area of 0.91 for round robin, ten-fold cross-validation and leave-one-out bootstrap. Likewise, all of the three validation methods have similar $_{0.90}$AUC. As expected, the resubstitution and classic bootstrap analysis have the highest AUC and $_{0.90}$AUC. Overall, there is no great difference in performance in AUC or $_{0.90}$AUC based on the method. The five methods do differ in the estimates of the standard deviation of the ROC measurements. The leave-one-out bootstrap appears to be the most rigorous and exhaustive evaluation, and also has the highest variance. It is likely that the high variance demonstrated by the bootstrap is reflective of the true variance that would be expected when evaluated on a new but similar data set, rather than using the more limited four other evaluation methods. For all five methods, the number of benign cases that could potentially be spared at 100% sensitivity on average (specificity) is very high. This means that benign cases could be spared from biopsy, while classifying all malignancies.

A comparable evaluation was carried out on subset B (defined in table 2), with similar results. These results are omitted for brevity.

3.2. Utilizing 'missing' information category

The leave-one-out bootstrap was applied to the evaluation of a single feature, in order to exemplify the potential information content in 'missing' information category. The individual feature used here as an example was hormone-therapy history under two conditions. Condition 1 was the classification of cases using the LRb as we have described in section 2, section B (*vide supra*). The LRb uses the distribution of 'missing' values in the H_0 and H_1 populations as information. In condition 2, the LRb was also used for classification, though no information about the distribution of missing values in the training set was used. For condition 2 the mean value (Beale and Little 1975) of the feature in the training set was substituted for the missing values in the testing set. The missing data method applied in condition 2 is a simple, often-utilized in practice method of dealing with missing data. The results for both conditions are as follows. Condition 1 had AUC = 0.56 ± 0.03 and PAI = 0.07 ± 0.01. Condition 2 had AUC = 0.53 ± 0.02 and PAI = 0.05 ± 0.004. Both AUC and

Table 4. Training and testing on various subsets of the data using the leave-one-out bootstrap. Three thousand bootstrap samples were used for each set-up, and thus each row represents an average of 3000 retraining runs of the classifier.

Set-up number	Description	Number of features	Training subset (cases)	Testing subset (cases)	AUC	$_{0.90}$AUC	Specificity at sensitivity of		
							95%	98%	100%
1	Train and test on all available cases, use all features	16	A (671)	A (671)	0.91 ± 0.02	0.60 ± 0.06	64%	47%	32%
2	Train and test on all cases, using max number of filled features	7	A_7 (671)	A_7 (671)	0.91 ± 0.02	0.57 ± 0.07	63%	45%	19%
3	Train and test on cases that have all feature values	16	B (225)	B (225)	0.91 ± 0.03	0.61 ± 0.13	66%	44%	44%
4	Train on cases that have at least 1 value missing, test on cases that have no values missing	16	C (446)	B (225)	0.91 ± 0.02	0.59 ± 0.08	62%	47%	39%
5	Train and test on cases that have at least 1 value missing	16	C (446)	C (446)	0.90 ± 0.02	0.58 ± 0.07	62%	50%	34%
6	Train on cases that have no values missing, test on cases that have at least one value missing	16	B (225)	C (446)	0.88 ± 0.02	0.45 ± 0.07	48%	30%	17%

PAI were better for condition 1, using information content in 'missing' information category. AUC was better in 83% of the bootstrap samples and PAI was better in 99% of the bootstrap samples. (Evaluation of FPFs at specific sensitivities is not appropriate, due to the small number of points on the ROC curve.) For hormone-therapy history, it was more advantageous to utilize the distribution information about the missing values, than to use the mean value to substitute for missing values. In effect, 'missing' information category contained some information about the likelihood of malignancy.

3.3. Performance of classifier using all features; effect of missing values

We carried out six different training set-ups for evaluating the performance with and without missing feature values. The case subsets used are as defined in table 2. The results for these set-ups are listed in table 4. Please note that 3000 bootstrap samples were used for each evaluation. Each bootstrap sample corresponds to a new retraining and testing of the classifier.

3.3.1. Comparison of set-ups 1 and 2. The first set-up consisted of training and testing the classifier on all 671 available cases with all available 16 features (data set A, table 2). This means that some of the cases had feature values missing. Set-up 2 involved training and testing

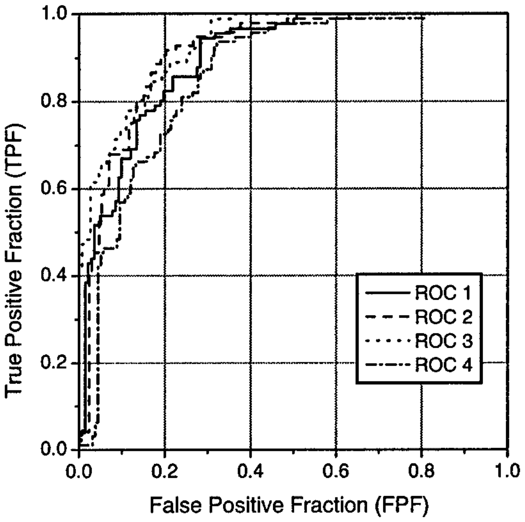

Figure 1. Four example ROC curves from the 3000 bootstrap evaluations of set-up 1. Figure 2 shows the accumulation of all 3000 ROC curves.

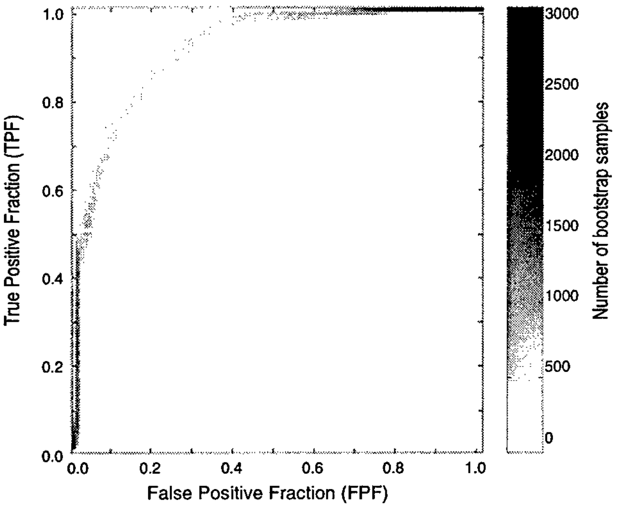

Figure 2. Cumulative ROC over 3000 bootstrap samples for set-up 1, using 671 mass biopsy cases and 16 features.

on all 671 available cases, but using only the features that had values for all cases (data set A_7, table 2). This means that in set-up 2 only 7 features for all cases were used. Leave-one-out bootstrap was used for training/testing both set-ups, and the case order from set-up 1 was also used in set-up 2 to allow pair-wise comparison of boot sample results.

Figure 1 shows four examples of ROC curves for four bootstrap samples for set-up 1. Figure 2 shows the accumulation of all 3000 ROC curves on an image. The darker the pixel

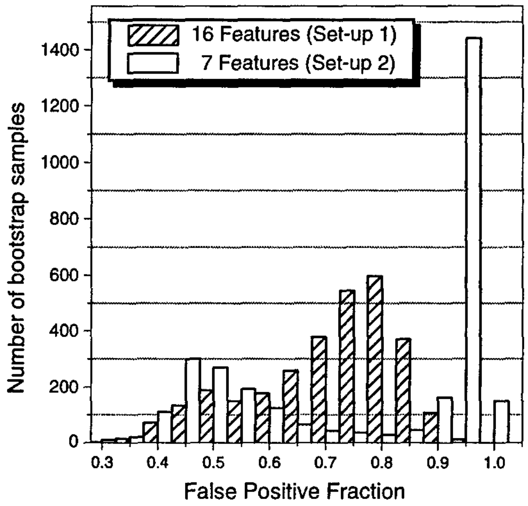

Figure 3. Histogram of false positive fractions at the 100% sensitivity level for the 3000 bootstrap runs using all 671 cases. Set-up 1 (crosshatch bars) represents the runs using all 16 features, even when some feature values were missing. Set-up 2 (white bars) represents the runs using the seven features that were present for all 671 cases. On average, set-up 1 has better specificity (lower false positive fractions) than set-up 2 at the 100% sensitivity level.

on the image, the higher the number of ROC curves that passed through that point. Only the actual discrete points were plotted on a 100 × 100 grid, with no interpolation between the points.

Both set-ups 1 and 2 had an AUC of 0.91 ± 0.02. The ${}_{0.90}$AUC was only slightly higher when using all 16 features (0.60 versus 0.57) rather than 7 features. However, there was a noticeable difference in performance at the 100% sensitivity level. The histogram of false positive fractions for the 100% sensitivity level (${}_{100\%}$FPF) is shown in figure 3. The two distributions are clearly different. As shown in figure 3, when using only seven features (set-up 2), the most often occurring ${}_{100\%}$FPF is 97%. This indicates that in many instances, only 3% of benign cases could potentially be spared from biopsy in set-up 2. For set-up 1, the most often occurring ${}_{100\%}$FPF is around 78%. In figure 4, the differences between the ${}_{100\%}$FPF of the paired bootstrap samples are plotted. Though no method is available to determine the statistical significance of this difference, it appears from figure 4 that it is more advantageous to use all information in the 16 features. Initially, it may have seemed that only the seven features are sufficient for great prediction and that we can exclude features that are missing some values. Including all the extra features did not significantly raise AUC or ${}_{0.90}$AUC. However, the inclusion of the extra features (that have some missing values) does raise the average specificity at high sensitivities, specifically at the 100% sensitivity level. The 100% sensitivity level is of importance in this application, since it allows to spare benign cases from biopsy while correctly identifying all the malignancies.

3.3.2. Comparison of set-ups 3 and 4. The third and fourth set-ups consisted of testing the classifier on subset B (table 2), 225 cases that had all the feature values. The difference between set-ups 3 and 4 was again the training data set. Set-up 3 was trained on cases that had all feature values (subset B), and set-up 4 was trained on cases that had at least one value missing (subset C, table 2). In order to better evaluate the effect of the training data set, the

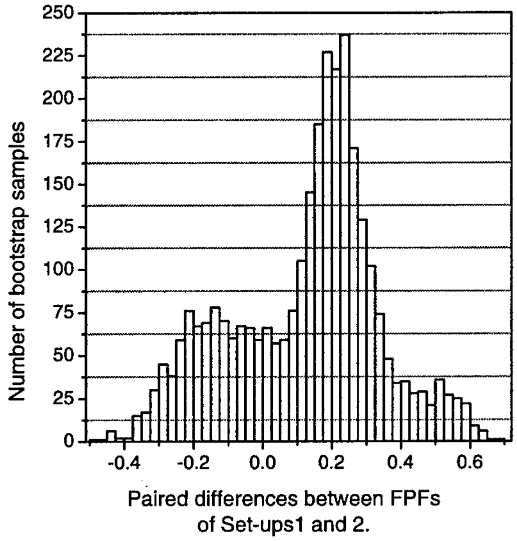

Figure 4. Histogram of paired differences between the false positive fractions (FPFs) of set-ups 1 and 2 at 100% sensitivity for 3000 bootstrap samples.

testing cases that were drawn randomly for set-up 3 were used in the same order for testing in set-up 4. There was no difference in the performance of set-ups 3 and 4: both had an AUC of 0.91, and similar $_{0.90}$AUCs (0.61 and 0.59). Furthermore, there was little difference in the FPFs at 100% sensitivity as shown in figure 5. Figure 5 shows that the two $_{100\%}$FPF distributions are very similar. In set-up 3, 44% of benign cases would be spared on average at the 100% sensitivity level. In set-up 4, 39% of benign cases would be spared on average. These results suggest that when the testing data set is not missing values, the training data set may or may not contain missing values. There will be no need to use or estimate data for missing values, since there are none in the testing set. Therefore, the performance will be influenced only by how well the training data represent the testing data, and not by how missing values are handled. In this comparison, both subsets B and C represent the testing data subset B equally well, with no difference in performance.

3.3.3. Comparison of set-ups 5 and 6. Set-ups 5 and 6 demonstrate the effect of training data when the classifier is tested on a data set with missing values (subset C). In set-up 5, leave-one-out bootstrap is used to train and test the classifier on cases that have at least one feature value missing (subset C). In set-up 6, the classifier is tested on C, but trained on subset B, the 225 cases that have all feature values. There was little difference in performance of set-ups 5 and 6. The AUCs are similar (0.90 ± 0.02 versus 0.88 ± 0.02), and the $_{0.90}$AUCs are similar (0.58 ± 0.07 versus 0.45 ± 0.07). However, for 94% of the bootstrap samples, the PAI was higher for set-up 5, and the AUC was higher for 83% of the samples for set-up 5. Figure 6 shows the distribution of FPFs at the 100% sensitivity level. The distribution of 100% FPFs for set-up 6 is slightly worse when compared to set-up 5. On average, set-up 5 has higher FPFs at the 100% sensitivity than set-up 6 for 77% of the bootstrap samples. This difference does suggest that when *testing* on a database with missing values, it is also beneficial to *train* on a database with missing feature values.

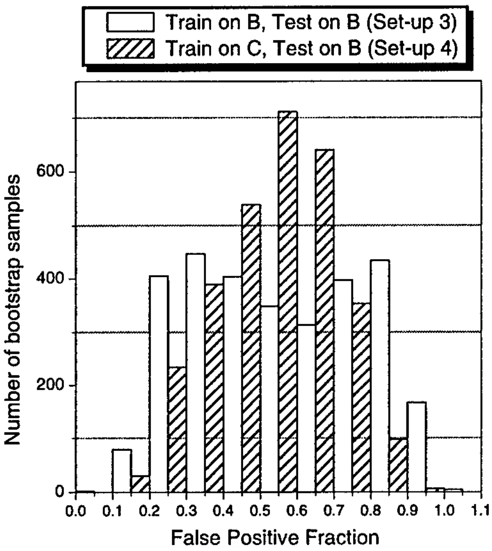

Figure 5. Histogram of false positive fractions at the 100% sensitivity level for the 3000 bootstrap runs for testing on subset B (225 cases with all 16 feature values). Set-up 3 was trained and tested on subset B using leave-one-out bootstrap. Set-up 4 was trained on subset C (446 cases that have at least one value missing), and tested on subset B. There is no apparent difference in the performance at the 100% sensitivity level.

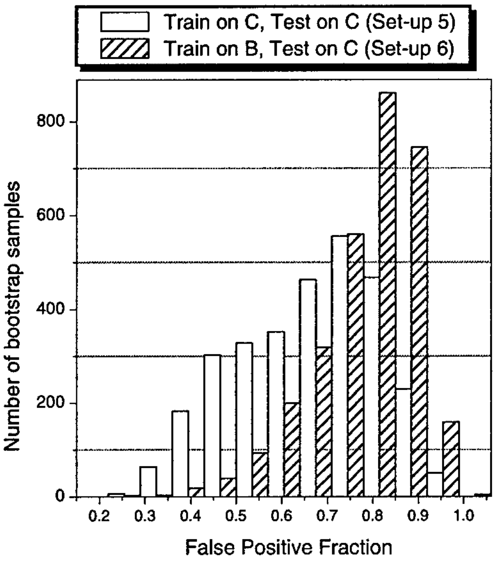

Figure 6. Histogram of false positive fractions at the 100% sensitivity level for the 3000 bootstrap runs for testing on subset C (446 cases that have at least one value missing). Set-up 5 was trained and tested on subset C using leave-one-out bootstrap. Set-up 6 was trained on subset B (225 cases with all 16 feature values), and tested on subset C. On average, set-up 5 has slightly better specificity (lower false positive fractions) than set-up 6 at the 100% sensitivity level.

4. Discussion

The need to reduce the number of biopsies performed on benign findings is a well-recognized issue in mammography. It is hoped that some of the issues faced by mammography can be alleviated by computer-aided diagnosis. In this study, we presented a highly specific computer classifier that could aid physicians by identifying benign lesions that should not be sent to biopsy. The classifier was developed on a medical database with some incomplete information.

Lack of values for certain cases in a database can prevent a classification algorithm from utilizing all the information that is present in the database. For example, in this study we had four choices in what data we chose for developing a classification tool. (1) Use all available cases, and use only the features that were recorded for all the cases. In this situation only seven features would be used. (2) Use all available features, and be limited to only the cases that have all features (225 cases). In either option 1 or 2, a very large portion of the available data is discarded. (3) We can compromise between options 1 and 2, and choose some portion of the cases and features that satisfies our need, resulting with a number of cases between 225 and 671, and the number of features between 7 and 16. Many studies choose this option as a solution to the missing data problem. We also have used this approach in previous studies (Floyd *et al* 2000, Bilska-Wolak and Floyd 2002a). However, there is no standard method for choosing these peculiar numbers and usually arbitrary empirical techniques are used. This method also discards some of the data that may contain useful information. (4) We can develop a classifier that can cope with missing values in input data. For this study, we chose option 4, and evaluated the classifier's performance with and without missing data.

For our classifier evaluation, we examined six different set-ups for training and testing on the given database. There are a lot of other set-ups/combinations that could have been applied in evaluating the effects of missing data and cases. These six set-ups were chosen to best represent possible effects on performance when faced with a database with some missing data. Other possible combinations would be interesting to examine, but would not be practical to include here in one study.

In the six set-ups, regardless of which data subset was used for training, the performance was almost identical in AUC. The AUC was around 0.90 for all six set-ups. This suggests that the lack of certain feature values did not substantially affect AUC. The performance was also very similar in ${}_{0.90}$AUC. The largest difference ${}_{0.90}$AUCs was evident between set-ups 5 and 6, suggesting that when testing on a database with missing values, it is beneficial to also train on a database with missing values. A difference in the performance of the six set-ups is also evident when we examine the distributions of FPFs at the 100% sensitivity level. It is clear from the distribution of FPFs of set-ups 1 and 2 (figures 3 and 4) that it would be preferable to use 16 features with some missing values, than only 7 features that had all values. Similarly from the distribution of FPFs, when testing on cases that are missing feature values, the classifier should also be trained on cases that are missing feature values for best performance. We would like to reiterate here that the inclusion of cases with missing data is not a trivial problem. Regardless of whether there is or not a statistical significance of the difference, the merits of the algorithm are scientifically interesting. The fact that there is no worsening of the results by the inclusion of cases with some missing data is in itself an important step. Full statistical validation of whether this difference is statistically significant will have to wait until a more sophisticated method is available. As evident in the figures, the results appear better than anything done previously.

The LRb's performance is not higher than previous classifiers' in terms of AUC or ${}_{0.90}$AUC. However, the LRb does exhibit better performance at high sensitivities than previous classifiers on the same data set, especially at the 100% sensitivity level. The 100% sensitivity level is very desirable in this application, since it allows us to spare benign cases from biopsy while correctly identifying all the malignancies. With the LRb trained/tested on our whole database of 671 cases, we can spare 32% of the benign cases from biopsy while correctly identifying 100% of the malignancies. For example, a previous case-based reasoning (CBR) classifier trained on almost the same case subset achieved an AUC of 0.91 ± 0.01 and PAI and 0.60 ± 0.05 (Bilska-Wolak and Floyd 2002b). This performance is similar to the LRb performance, but the CBR was not able to spare any benign cases at the 100% sensitivity level. A neural network classifier trained on almost exactly same data also achieved an AUC of 0.93 ± 0.01 and PAI of 0.62 ± 0.05 (Markey *et al* 2002). In the same study, linear discriminant analysis achieved an AUC of 0.91 ± 0.01 and PAI of 0.61 ± 0.04. The AUC and PAI performances are similar to the LRb, while the LRb is a much simpler classifier. No values for specificities at sensitivities were reported for the ANN or discriminant analysis in this study. The LRb appears able to achieve similar AUC and ${}_{0.90}$AUC, but also high specificity at 100% sensitivity, in contrast to other classifiers previously utilized for this problem.

The LRb is well suited for training/testing on cases with missing values because it does not directly depend on the rankings/number scale assigned to feature values. Several of the commonly used classifiers will have difficulty handling this situation. Ideally, one should collect all the needed data for all cases. However, situations may and do exist when all data are not available. In such situations, it appears better to use an informative feature with missing values for some cases, than not to use the feature at all. A classifier such as the LRb is well suited for the classification of data with missing feature values.

4.1. Limitations of this study

It is not possible to interpret our results without noting that relatively few feature values (1.3 ± 1.1) were missing in our database. Also, most of the key features were recorded for the majority of cases. For example, all cases had mass margin value recorded. There will exist a point when the lack of data corresponds to such lack of information that no feasible prediction can be accomplished. However, when only some cases are missing few of the features, it is possible to utilize the full database and maximize the number of features and cases. The application of LRb allowed the usage of all the information present in the available database.

This study did not address the issue of artificially removing data points and evaluating the performance with continuously decreasing size of data. Rather, it focused on a real-life situation one might face when obtaining data for classifier development. No imputation of missing data was performed to fill in the missing data values. It is possible that imputation would also improve performance especially in situations where no cases with missing data are available for training. Also, there are many other possible set-ups/combinations, and we examined just six which were considered most representative.

For breast biopsy prediction, the most desirable operating point is 100% sensitivity. We would like to operate as close as possible to this point so that we can correctly classify all malignancies. However, analysis at very high sensitivities should be regarded cautiously due to difficult statistical character of distribution extremes. The use of the bootstrap allowed a more thorough investigation of the behaviour of the classifier over a wide range of data. While our leave-one-out bootstrap visual analysis of specificities at 100% sensitivity can be considered statistically incomplete, the odds of operating at 100% sensitivity with good results will be higher when we formulate such higher goals. Classifier performance at 100% sensitivity is a clinically significant measure of performance for breast biopsy classification.

No feature selection was applied in this study. It is possible that feature selection might improve the performance of the classifier. It is also possible that no improvement will be observed, or simply that identical performance will be achieved with fewer features. In this study, we demonstrated that excellent performance for a LRb classifier can be achieved without feature selection. Lack of feature selection decreases the potential for over-training the system. Nonetheless, feature selection may be an issue worth addressing in the future.

5. Conclusions

The implementation of the LRb allowed us to utilize all the information present in the medical database. Compared with other efforts, this enhanced the performance at high sensitivities. The high specificity at 100% sensitivity may be explained by the use of all available information by the LRb. No feature data and no cases are discarded due to missing values. This means that more information is available to accurately estimate the true feature distributions. Also it was established that when testing on cases with missing feature values, the database should be trained on cases with missing values for best performance.

The LRb performance was exemplary for the task of biopsy classification. Without computing the ROC curve for the physicians' performance, we can state that this classifier has achieved better performance than an individual physician. Note that the classifier is trained on the collective knowledge of numerous physicians. Theoretically, each physician operated at the 100% sensitivity point, and conservative behaviour resulted in the misclassification of many benign cases. The LRb was able to maintain 100% sensitivity, while correctly identifying on average 32% of benign cases that should be spared from biopsy. This suggests that the classifier could be a valuable tool for individual physicians, by decreasing the number of benign cases sent to biopsy.

Acknowledgments

This work was supported in part by USAMRMC grant number DAMD17-03-1-0015 awarded by the US Army; USPHS grant number R21-CA81309, awarded by the National Cancer Institute; USAMRMC grant number DAMD17-99-1-9174 awarded by the US Army, and Susan G Komen Breast Cancer Foundation grant number BCTR2000730A.

References

Barrett H H, Abbey C K and Clarkson E 1998 Some unlikely properties of the likelihood ratio and its logarithm *SPIE Med. Imaging: Image Perception* **3340** 65–77

Beale E M L and Little R J A 1975 Missing values in multivariate analysis *J. R. Stat. Soc.* B **37** 129–45

Bilska-Wolak A O and Floyd C E Jr 2002a Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with BI-RADS(TM) lexicon *Med. Phys.* **29** 2090–100

Bilska-Wolak A O and Floyd C E Jr 2002b Breast biopsy prediction using a case-based reasoning classifier for masses versus calcifications *SPIE Medical Imaging; 2002 (San Diego, CA)* ed Sonka M and Fitzpatrick J M pp 661–5

BI-RADS 1998 *American College of Radiology Breast Imaging—Reporting and Data System (BI-RADS)* 3rd edn (Reston, VA: American College of Radiology)

Chang Y H *et al* 2001 Knowledge-based computer-aided detection of masses on digitized mammograms: a preliminary assessment *Med. Phys.* **28** 455–61

Dixon J M and John T G 1992 Morbidity after breast biopsy for benign disease in a screened population *Lancet* **1** 128

Doi K, MacMahon H, Katsuragawa S, Nishikawa R M and Jiang Y 1999 Computer-aided diagnosis in radiology: potential and pitfalls *Eur. J. Radiol.* **31** 97–109

Efron B and Tibshirani R J 1993 *An Introduction to the Bootstrap* (New York: Chapman and Hall)

Efron B and Tibshirani R J 1997 Improvements on cross-validation: the 0.632+ bootstrap method *J. Am. Stat. Assoc.* **92** 548–60

Egan J P 1975 *Signal Detection Theory and ROC Analysis* (New York: Academic)

Floyd C E Jr, Lo J Y and Tourassi G D 2000 Cased-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions *Am. J. Roentgenol.* **175** 1347–52

Gavrielides M A, Lo J Y, Vargas-Voracek R and Floyd C E Jr 2000 Segmentation of suspicious clustered microcalcifications in mammograms *Med. Phys.* **27** 13–22

Hadjiiski L, Sahiner B, Chan H, Petrick N, Helvie M and Gurcan M 2001 Analysis of temporal changes of mammographic features: computer-aided classification of malignant and benign breast masses *Med. Phys.* **28** 2309–17

Hand D, Mannila H and Smyth P 2001 *Principles of Data Mining* (Cambridge: MIT Press)

Helvie M A, Ikeda D M and Adler D D 1991 Localization and needle aspiration of breast lesions: complications in 370 cases *Am. J. Roentgenol.* **157** 711–4

Jain A K, Dubes R C and Chen C-C 1987 Bootstrap techniques for error estimation *IEEE Trans. Pattern. Anal. Mach. Intell.* **9** 628–33

Jiang Y, Metz C E and Nishikawa R M 1996 A receiver operating characteristic partial area index for highly sensitive diagnostic tests *Radiology* **201** 745–50

Jiang Y, Nishikawa R M, Schmidt R A, Toledano A Y and Doi K 2001 Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications *Radiology* **220** 787–94

Kaye M D, Vicinanza-Adami M L and Sullivan M L 1994 Mammographic findings after stereotaxic biopsy of the breast performed with large-core needles *Radiology* **192** 149–51

Kohavi R 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection *Int. Joint Conf. on Artificial Intelligence* (San Francisco, CA: Morgan Kaufman) pp 1137–43

Kopans D B 1992 The positive predictive value of mammography *Am. J. Roentgenol.* **158** 521–6

Kittler J, Hatef M, Duin R P W and Matas J 1998 On combining classifiers *IEEE Trans. Pattern. Anal. Mach. Intell.* **20** 226–39

Kupinski M A, Edwards D C, Giger M L and Metz C E 2001 Ideal observer approximation using Bayesian classification neural networks *IEEE Trans. Med. Imaging* **20** 886–99

Little R J A and Rubin D B 2002 *Statistical Analysis with Missing Data* 2nd edn (Hoboken, NJ: Wiley)

Lo J Y, Baker J A, Kornguth P J and Floyd C E Jr 1999 Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks *Acad. Radiol.* **6** 10–5

Markey M K, Lo J Y and Floyd C E Jr 2002 Differences between computer-aided diagnosis of breast masses and that of calcifications *Radiology* **223** 489–93

McDonough R N and Whalen A D 1995 *Detection of Signals in Noise* (San Diego: Academic)

Meyer J E, Eberlein T J, Stomper P C and Sonnefeld R R 1990 Biopsy of occult breast lesions: analysis of 1261 abnormalities *J. Am. Med. Assoc.* **263** 2341–3

Michaelson J *et al* 2002 The patterns of breast cancer screening utilization and its consequences *Cancer* **94** 37–43

Miller A B, Baines C J, To T and Wall C 1992 Canadian national breast screening study: 1. Breast cancer detection and death rates among women aged 40 to 49 years *Can. Med. Assoc. J.* **147** 1459–76

Morimoto T, Sasa M, Yamaguchi T, Kondo H, Akaiwa H and Sagara Y 2000 Breast cancer screening by mammography in women aged under 50 years in Japan *Anticancer Res.* **20** 3689–94

Population by Age, Sex, Race and Hispanic or Latino Origin: 2000 (PHC-T-9): US Census Bureau, Population Division; 2000 October 3 2001

Scott D W 1979 On optimal and data-based histograms *Biometrica* **66** 605–10

Scott D W 1992 *Multivariate Density Estimation: Theory, Practice, and Visualization* (New York: Wiley)

Swensson R G, King J L, Good W F and Gur D 2000 Observer variation and the performance accuracy gained by averaging ratings of abnormality *Med. Phys.* **27** 1920–33

Tabar L, Vitak B, Chen H H, Yen M F, Duffy S W and Smith R A 2001 Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality *Cancer* **91** 1724–31

Tourassi G D, Markey M K, Lo J Y and Floyd C E Jr 2001 A neural network approach to breast cancer diagnosis as a constraint satisfaction problem *Med. Phys.* **28** 804–11

VanTrees H L 1968 *Detection, Estimation, and Modulation Theory (Part I)* (New York: Wiley)

Veldkamp W J H, Karssemeijer N, Otten J D M and Hendriks J H C L 2000 Automated classification of clustered microcalcifications into malignant and benign types *Med. Phys.* **27** 2600–8

Wand M P 1995 Data-based choice of histogram bin width *Australian Graduate School of Management Working Paper Series No. 95-011* University of New South Wales

Zheng B, Chang Y, Good W F and Gur D 2001 Performance gain computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering *Med. Phys.* **28** 2302–8

Zhou X-H, Obuchowski N A and McClish D K 2002 *Statistical Methods in Diagnostic Medicine* (New York: Wiley)

Appendix C

A.O. Bilska-Wolak, C.E. Floyd Jr., Joseph Y. Lo, "Computer Aid for Decision to Biopsy Breast Lesions: Pre-clinical Performance Evaluation," (in submission).

Computer Aid for Decision to Biopsy Breast Lesions: Pre-clinical Performance Evaluation

Abstract

OBJECTIVE: The purpose of this study was to validate the performance of a previously developed computer-aid for breast mass classification on a new independent database of 151 cases that were not used for algorithm development.

MATERIALS AND METHODS: A computer-aid (classifier) based on the likelihood ratio (LRb) was previously developed on a database of 670 mass cases. The 670 cases (245 malignant) from one medical institution were described using 16 features from the BI-RADS^{TM} lexicon and patient history findings. Continued data collection yielded additional 151 (43 malignant) cases that were previously unseen by the classifier. These new cases were examined by the developed classifier. Performance evaluation methods included Receiver Operating Characteristic (ROC), round-robin, and leave-one-out bootstrap sampling.

RESULTS: The performance of the classifier on the training data yielded an average ROC area of 0.90+/- 0.02, and partial ROC area ($_{0.90}$AUC) of 0.60+/-0.06. The exact non-parametric performance on the independent set of 151 cases yielded a ROC area of 0.88 and $_{0.90}$AUC of 0.57. Using a 100% sensitivity cutoff threshold established on the training data, the classifier was able to correctly identify 100% of the malignant lesions in the new independent set, while potentially obviating 26% of the biopsies performed on benign lesions.

CONCLUSION: In this pre-clinical evaluation, the LRb classifier performed equally well on the new independent data that was not used for classifier development. The LRb classifier performance compared favorably with an artificial neural network. The LRb classifier shows promise as a potential aid in reducing the number of biopsies performed on benign lesions.

Introduction

X-ray mammography, while highly sensitive to mammographic abnormalities, has low specificity to benign lesions. The low specificity to benign lesions is due mainly to similar radiographic appearance of benign and malignant lesions, and the physicians' cautious recommendations. Consequently as many as 65-85%[1-3] of breast biopsies are performed on benign lesions. The potentially inordinate biopsy of benign lesions raises emotional, physical and financial burdens to the patients and clinic. [4-7] Overall, excessive biopsy reduces the efficacy of mammographic screening.

The diagnostic performance for suspicious lesions in mammography may be enhanced by computer decision aids that could supply additional information about the likelihood of malignancy. Such computer-aids represent a low-cost, non-invasive accessory to diagnosis by providing a swift second opinion that can be highly accurate. Computer-aids for mammography have been applied to the task of detection of mammographic lesions[8, 9] and to classification of suspicious lesions.[10, 11] In this work, we concentrate on the classification of mammographic mass lesions.

While a great amount of research has been performed in developing and testing computer aids for mammography during the last three decades, only recently has research appeared in which extensively trained CAD tools are tested on completely new data. By "new" we mean cases that have not been available for classifier training and development, and are thus "unknown" to the classifier. Such evaluation on new data is referred to as classifier validation. Although classifier *training* methods, such as Round-Robin and cross-validation, use portions of data at a time for development, they do not guarantee unbiased future performance estimation. This occurs since all available training data is eventually used for classifier development and the classifier has some "knowledge" of all the cases. Therefore, classifier validation is crucial in determining whether the trained classifier is generalizable to new unknown data and thus suitable for actual clinical application.

Validation studies of CAD classifiers for breast cancer show encouraging results for the future of CAD in mammography. Huo et.al.[12] evaluated a CAD system for the classification of mammographic masses on a 110 case database, achieving Az values (parametric area under the binormal ROC) of 0.82. Chang et.al.[13] assessed a CAD system for the detection of mirocalcification clusters on 386 new digitized mammogram images. They achieved 89.5% sensitivity at an average false-positive detection rate of 0.39 per image. Tourassi et.al. [14] performed the validation of a constraint satisfaction neural network on 1030 breast biopsy cases. Using mammographic and clinical findings, an Az of 0.81+/-0.02 was achieved for the biopsy cases. These studies have demonstrated that it is possible to develop a computer aid that is generalizable to new unknown cases, and thus promising for potential clinical application.

In our previous study, we developed a likelihood ratio based (LRb) classifier for breast biopsy prediction.[15] The classifier was developed on 671 breast mass biopsy cases. Suspicious mammographic mass lesions were described using BI-RADS^{TM} lexicon[16] and patient history findings. In the clinical environment, these findings would be entered by a clinician into the classifier. Based on the findings, the classifier would predict the likelihood of malignancy and suggest either biopsy or short-term follow-up. Since the BI-RADS^{TM} lexicon was developed to standardize mammography reporting, it is conceivable that the model should perform well regardless of the clinician that makes the BI-RADS^{TM} assessments.

Very high sensitivity is needed for the possible clinical acceptance of a computer tool that classifies lesions that have already been detected. It is considered a worse error in judgment to misclassify a detected lesion as benign, when in fact it is malignant. High sensitivity presents a great challenge to any computer-aid that attempts to assist in the lesion classification task.

We report on the performance of the classifier on the 670 cases (due to inconsistencies, one case was subsequently removed from the database for this study). In the present study, we present the performance of the classifier on a new independent data set of 151 cases that was not included in the classifier development. We document that the classifier was robust and its performance remained high on the new data.

Methods

Description of databases

The original data, collected at one medical center between 1988 and 2000, consisted of 670 biopsy cases that contained a suspicious mass on a mammogram. These mass cases were defined as cases that had a mass lesion and any other findings. These other findings included, for example, microcalcifications and architectural distortions. The data were collected as part of standard clinical practice, with approval from appropriate institutional review boards. All physicians were dedicated breast imaging radiologists. Since all the cases were sent to biopsy, the biopsy outcome was available from the histopathological analysis. Of the 670 cases, 244 (36%) proved malignant at biopsy, while the rest - 426 (64%) - were benign.

Each suspicious mammographic lesion was described by a dedicated breast-imaging radiologist using the BI-RADS^{TM} lexicon. Each case in the database was thus represented using sixteen features based on the BI-RADS^{TM} lexicon and patient history. These sixteen features included mass margin, mass shape, mass density, mass size, calcification morphology, calcification distribution, calcification number, associated findings, special cases, quadrant location, change from prior mammogram, breast side, architectural distortion as main finding, hormone use, menopausal status and patient age. For the 670 cases, 1.3 +/- 1.1 feature values were missing (not recorded) per case. However, no cases were excluded on the basis of missing features, and all 670 cases even with some missing values were utilized for the algorithm/classifier development. Table 1 lists the characteristics of the training/testing data sets.

Ongoing data collection at the medical institution yielded another 151 biopsy cases. While this case subset became available during the course of our research, it was not included in classifier training and development. This subset contained 108 benign, 43 malignant (151 total) mass cases that were collected similarly to the original set. However, all of the new validation cases were missing two feature values: menopausal status and hormone use. This information had not been collected and was not available for the present evaluation. As a result, 3.2 +/- 1.0 feature values were missing per validation case. The case composition of the sets may represent the clinical situation, in which past cases are used to determine the outcome of more recent cases with differing characteristics. If the two sets were identical in characteristics, it would not be necessary to perform a validation test. The characteristics of the 151 case independent set are listed in Table 1.

Computer Classifier

The computer classifier developed here for breast mass classification was based on the likelihood ratio (ideal observer) paradigm. The likelihood ratio provides the optimal classifier for the binary outcome situation (malignant/benign), given the true feature distributions.[17-19] The true feature distributions for the categorical BI-RADS^{TM} features were estimated using a histogram approach. This means that the features and thus classifier were not dependent on values assigned to the feature findings. Each feature distribution was estimated using all available cases, and then the outputs from each distribution were merged to give the final classifier output. The classifier was developed and trained on the previously available 670 mass Duke cases.

Actual clinical application of a classifier requires establishing a sensitivity threshold to decide which cases are malignant and which cases are benign. In practice, this is accomplished by establishing a cutoff threshold from the ROC curve (*vide infra* Classifier Evaluation Methods) of the training data. We evaluated four sensitivity thresholds (100%, 99%, 98%, and 95%) that might be clinically acceptable for biopsy classification. For each sensitivity level, two different methods were utilized to determine their ROC thresholds. Since non-parametric ROC analysis was used, the thresholds were determined based on the ROC curves of the training/testing data set for both methods. The first method involved the establishment of thresholds directly from the Round Robin (RR) ROC curve of the 670 training cases, and we will refer to it as *threshold method RR.* The second method, named *threshold method BB,* involved the establishment of most often occurring thresholds from the bootstrap evaluation of the 670 training cases. For a given sensitivity level (eg. 100%), the distribution of 3000 thresholds from each bootstrap run was examined. The peak of this distribution was established as the most likely and consistent threshold for the given sensitivity level.

Classifier Evaluation Methods

Receiver Operating Characteristic (ROC)[20, 21] analysis was used to evaluate the classifier. The ROC curve illustrates the performance of a classifier over all sensitivity and specificity levels. A succinct measure of performance is the area under the ROC curve (AUC). [20] The AUC can be described as the average specificity over all sensitivities. The

AUC can range from 0.5 (chance performance) to 1.0 (perfect performance). Other measures of performance relating to the ROC curve include partial area index[22] ($_{0.90}$AUC) above the sensitivity of 90%, which represents the average performance of the classifier at sensitivities from 90% to 100%. The partial area is a more clinically meaningful measure of performance when high sensitivity is essential. Other measures of performance may include points on the ROC curve corresponding to specific sensitivity and specificity. The sensitivity at a fixed specificity is often preferable to the AUC when evaluating a test for a particular application.[23] Since in this application we are interested in correctly diagnosing malignancies while decreasing the number of biopsies performed on benign findings, we are interested in very high sensitivity and high specificity. Additional measures of performance included specificity at 95% sensitivity, specificity at 98% sensitivity, and specificity at 100% sensitivity. Although specificity at a high sensitivity (such as 100%) may seem like a very restrictive goal, other studies have also used this operating point for specific applications.[24] In this application, specificity at 100% sensitivity indicates how many benign cases could be potentially spared from biopsy while correctly diagnosing all malignancies.

Training and Testing. Leave-one-out bootstrap[25, 26] was used for classifier training and testing. Leave-one-out bootstrap (also referred to ε_0) is a subcomponent of the 0.632 bootstrap,[26, 27] without the error correction. This means that theoretically, the leave-one-out bootstrap gives pessimistic estimates of performance. For the training/testing stage, 3000 bootstrap samples were drawn with replacement from the original case set. Each bootstrap sample was used to train the classifier, and the cases not in the bootstrap sample were used to test the classifier. This resulted in 3000 estimates of performance, from which the mean performance (ROC measurements) and standard deviations were computed. In effect, the classifier was trained and tested 3000 times, each time on a different subset of the training /testing data.

Independent Validation. To simulate the performance of the optimized LRb classifier in a potential clinical situation, the classifier was applied to analyze the independent case set. The performance on the independent case set was calculated from the validation ROC curve. The thresholds that had been established on the training data were applied directly to the validation ROC curve. This yielded a binary cutoff for establishing a clinical recommendation: proceed with biopsy or with short-term follow-up. The performance at the thresholds (95-100% sensitivity) was examined in terms of numbers of cases that would be correctly diagnosed as benign and malignant. This corresponds to the exact maximum performance one would expect if the classifier was used in the clinic.

Results

Results of Training and Testing

The results of the training and classifier development are listed in Table 2 (evaluations 1 & 2). The leave-one-out bootstrap evaluation represents the average of the 3000 training-runs of the classifier. Each training-run was performed with a slightly different subset of the 670 cases. The average AUC was 0.90+/- 0.02, while the partial area $_{0.90}$AUC was 0.60 +/-0.06. On average, the classifier achieved 32% specificity at 100% sensitivity. A simple Round Robin evaluation was also carried out and the values for this evaluation are listed in Table 1. Figure 2 shows the resulting Round Robin (RR) curve for the training/testing of 670 cases (curve A).

Figure 3 shows the distribution of cutoff thresholds established for each of the four sensitivities. These threshold distributions were established from the bootstrap evaluation. The peak of each distribution was chosen as the most likely threshold for establishing the given sensitivity on the training data. These peaks are the BB method thresholds.

Results of Independent Validation

Figure 2 shows the curve for the independently evaluated 151 cases (curve B). The exact ROC results for the independent evaluation (Table 2) are as follows: AUC = 0.88, and $_{0.90}$AUC = 0.57. This result is similar to the Round Robin training evaluation of curve A (AUC=0.90, $_{0.90}$AUC =0.61, Table 2). Since to our knowledge there is no suitable method to compare these two results, we cannot discuss the statistical significance of the non-parametric difference. As shown in Figure 2, the two curves and their AUC values are very close. Considering that the classifier had no prior knowledge of any kind of these 151 cases, the results of the independent performance on the new data set are very promising for potential future clinical application.

A Round Robin evaluation of the independent data was also performed. The results of this evaluation are listed in Table 2. The Round Robin performance of the independent set was lower and had greater standard deviation than that of the validation. This suggests that the independent data set did not contain enough information to reliably train the classifier, even for Round Robin testing on itself.

Table 3 shows the performance on the validation set using the four cutoff thresholds established from the training/testing data. The most conservative threshold with impressive performance was 100% sensitivity. For the BB method, applying the 100% sensitivity threshold to the new data set actually yielded 98% sensitivity (misclassification of 1 malignancy) and specificity of 32%. This means that 32% of the benign lesions would be spared from biopsy. On the other hand, using the RR method threshold, we achieved 100% sensitivity and 26% specificity. This means that no malignancies would be misclassified, while obviating 26% of the biopsies on benign lesions. For the 100% sensitivity level, the RR method yielded higher resulting sensitivity than the BB method. The 100% - 95% sensitivity thresholds are also plotted on the validation curve in Figure 4. The classifier and both threshold methods produced encouraging results.

For the sake of completeness, we performed an analysis to evaluate the effect of bootstrap thresholds. Since the bootstrap evaluation produces a distribution of values for each measurement of interest (such as specificity at 98% sensitivity, Figure 3), the 95% confidence interval thresholds and the resulting sensitivity were examined (Table 4). For example, when the lower threshold established on the bootstrap evaluation was applied to the round robin curve of the 670 cases, it yielded 97% sensitivity. The upper threshold yielded 100% sensitivity. These endpoint thresholds were also applied to the ROC curve of the independent validation set. The lower threshold corresponding to 100% sensitivity corresponded to 91% sensitivity on the validation ROC curve. The upper threshold corresponded to 100% sensitivity on the validation ROC curve. The results for other thresholds/sensitivities are listed in Table 4.

Performance Comparison to Artificial Neural Network

An independent evaluation was also carried out using the same data sets and an artificial neural network (ANN[10]). The ANN was a three layer, feed forward and error-back propagation network. The ANN was trained on the 670 cases using 10-fold cross-validation. All network parameters were established empirically and were fixed after the training. The training yielded AUC of 0.93 +/- 0.01, and $_{0.90}$AUC of 0.64. At 100% sensitivity, the ANN had 0% specificity. At 98% sensitivity, the ANN had 41% specificity. The AUC and $_{0.90}$AUC were higher for the ANN than the LRb for the training stage. However, the results for the ANN are lower than the LRb at the 100% sensitivity level.

The trained ANN was then evaluated on the independent set of 151 cases. This validation test yielded an ROC area of 0.83+/-0.04, and $_{0.90}$AUC of 0.23+/0.10. Applying the 98% training threshold to the independent ROC yielded 91% resultant sensitivity, and 41% specificity. The results of the independent validation of the ANN were lower than the results of the LRb on this new independent data set. The differences were almost significant for AUC (p=0.058) and significant for $_{0.90}$AUC (p=0.002). Although the ANN was trained in a cross-validated manner with care to prevent over-training, the ANN still evidently overfit the training data. The ANN produced poorer results than the LRb on the new independent data set.

Discussion

The large number of biopsies performed on benign lesions is a well-recognized issue of mammography. One potential solution to this problem incorporates the use of computer-aided diagnosis tools that could offer a second opinion to a physician about a suspicious mammographic lesion. To standardize the reporting on suspicious mammographic lesions, the BI-RADS^{TM} lexicon was developed. In this and previous studies,[28, 29] we have shown that it is possible to utilize this standard lexicon to develop a computer-aid (LRb) that can classify suspicious mammographic lesions. The computer-aid helps to identify lesions that are likely benign and should not be sent to biopsy, which would result in a reduction in the number of benign lesions sent to biopsy. The high sensitivity of the LRb classifier prevents the misclassification of malignant lesions. The reduction in biopsies performed on benign lesions would reduce the stress and risks to patients, increase the effectiveness of mammographic screening and decrease costs.

This study performed the first step in evaluating the potential application of the LRb computer-aid in the clinic. This pre-clinical evaluation involved the analysis of cases previously unseen by the classifier. The cases represented a real-life situation, being recent and containing missing values. (The next step, not addressed in this study, would be evaluating the benefit to a physician utilizing this classifier.) In order to evaluate the LRb classifier in this pre-clinical manner, distinct performance measures were obtained from the ROC. These included the high sensitivity area of the ROC curve, as well as specificities at high sensitivities. We even analyzed the specificity at 100% sensitivity, which determines how many benign cases would be potentially spared from biopsy, while detecting 100% of malignant lesions. This measure is the most conservative performance threshold for our application.

The performance of the classifier, using even the most conservative performance measures, was commendable on a new independent data set previously unseen by the classifier. Using thresholds established on the training data, the classifier was able to correctly identify 26% of the benign lesions, while maintaining 100% sensitivity to malignant lesions. This represents an improvement over individual physician's performance, as all cases in the data set were originally sent to biopsy. This improvement can be accomplished by utilizing the collective knowledge of the physicians by the LRb classifier.

Another classifier, an ANN developed for this problem, had lower validation performance on the new data set than the LRb. There are several reasons that could explain the ANN's poorer performance with respect to the LRb. The development time invested into training the original network was significantly smaller than that time invested in developing the LRb. The ANN analyzed the cases using the same feature-to-number encoding scheme as in prior research, and was probably dependent on the numbers assigned to feature values. Furthermore, the ANN also utilized the information about missing feature values in simple manner. This may have had a detrimental effect on the ANN. However, training with fewer features (and thus much fewer missing values) had little effect on the ANN performance. This suggests that the ANN utilized only a portion of the information contained in all features, in contrast to the LRb. Using more hidden nodes did not improve ANN's performance. All these reasons could have contributed to the ANN's lower performance as compared to the LRb.

Several characteristics of the LRb classifier make it an attractive tool for the classification of mammographic lesions using BI-RADS^{TM} lexicon. The LRb utilizes the collective knowledge of numerous physicians to make a recommendation on a new case. The LRb can perform classification even when cases are missing some feature values. Missing feature values are a common real-life problem. Furthermore, the training/testing and validation data sets had differing characteristics (such as distribution of missing features) yet the classifier was able to perform commendably on the new data. The LRb appears more immune to over-training than an ANN, and is able to generalize well to previously unseen cases. The classifier is also computationally fast and simple, especially in the trained/developed state. All of these characteristics make the LRb a useful tool that could decrease the number of biopsies performed on benign lesions without compromising the classification of malignant lesions.

Conclusion

In this study, we evaluated the potential clinical application of the previously-developed LRb classifier for breast biopsy prediction. The classifier, developed on a database of 670 cases, was applied here to a new independent set of 151 biopsy cases. The classifier performed equally well on this new independent data that was not used for classifier development. This can be considered a successful pre-clinical evaluation of the classifier. The LRb classifier shows promise as a potential aid in reducing the number of biopsies performed on benign lesions.

Tables

Table 1: Characteristics of the training/testing set and the validation set.

Characteristic	Training/Testing Set	Validation Set
Number of cases	670	151
Number of malignant cases	244 (36%)	43 (28%)
Number of benign cases	426 (64%)	108 (72%)
Average age (age range)	56 years (24-87)	54 years (22 - 90)
Number of cases with microcalcifications	79 (12%)	13 (9%)
Number of architectural distortions	2	4
Number of missing feature values per case	1.3+/-1.1	3.2+/-1.0

Table 2: ROC results for the A) LRb trained and tested on the 670 mass cases, and B) for the LRb trained on the 670 cases and tested on an independent dataset of 151 cases.

					Specificity at Sensitivity of		
	Experiment Type (Number of Cases)	Evaluation Type	AUC	0.90AUC	95%	98%	100%
1	Training & Testing (670)	Leave-one-out bootstrap	0.90 +/- 0.02	0.60 +/- 0.06	64%	46%	32%
2	Training & Testing (670)	Round Robin	0.90 +/- 0.01	0.61+/- 0.04	66%	52%	25%
3	Validation (151)	Independent evaluation on 151 cases*	0.88	0.57	69%	57%	29%
4	Test of Validation Set (151)	Round Robin on 151 cases	0.86+/- 0.04	0.44+/- 0.17	48%	19%	17%

* the error bars are not included since the performance measure represents the exact values that would result when the developed classifier is applied to the new data

Table 3: Performance of Threshold Methods RR and BB at 100% - 95% sensitivity levels on the 151 independent cases.

ROC 151 Original Sensitivity	ROC 151 Original Specificity (maximum achievable)	BB Method Resulting Sensitivity	BB Method Resulting Specificity	RR Method Resulting Sensitivity	RR Method Resulting Specificity
100%	29%	98%	32%	100%	26%
99%	29%	98%	32%	98%	44%
98%	57%	95%	64%	95%	66%
95%	69%	88%	71%	81%	75%

Table 4: Bootstrap range in sensitivities applied to the Round Robin evaluation of the training/testing set and to the independent validation.

Sensitivity on Bootstrap Sample	**Range in Sensitivity for 95% of Bootstrap samples applied to Round Robin evaluation of training/testing set**	**Range in Sensitivity as applied to ROC curve of 151 cases**
100%	97% - 100% (3%)	91% - 100% (9%)
99%	96% - 100% (4%)	88% - 100% (12%)
98%	94% - 100% (6%)	79% - 98% (18%)
95%	88% - 98% (10%)	77% - 95% (19%)

Figures

Figure 1: Diagram of the trained classifier with inputs and outputs.

Figure 2: ROC curves for the A) LRb trained and tested on the 670 mass cases using Round Robin, and B) for the LRb trained on the 670 cases and tested on an independent dataset of 151 cases.

Figure 3: The distributions of thresholds (Beta) at A) 95%, B) 98%, C) 99%, and D) 100% sensitivity levels. These distributions were obtained from the bootstrap evaluation on 670 cases. The peak of each distribution was used to determine the most likely threshold for the given sensitivity. These cutoff thresholds are referred to as the BB method thresholds.

Figure 4: The performance of thresholds applied to the validation ROC curve of 151 cases. The thresholds were determined from the training on 670 cases using BB and RR methods. Next to each threshold is the original training sensitivity. Note that the first drop in sensitivity on the ROC curve (to 98%) corresponds to a misclassification of one malignant case in the independent set.

REFERENCES

[1] D. B. Kopans, "The positive predictive value of mammography," AJR **158**, 521-526 (1992).

[2] J. E. Meyer, T. J. Eberlein, P. C. Stomper, and R. R. Sonnefeld, "Biopsy of occult breast lesions: Analysis of 1261 abnormalities," J. Am. Med. Assoc. **263**, 2341-43 (1990).

[3] S. Ciatto, L. Cataliotti, and V. Distante, "Nonpalpable lesions detected with mammography: review of 512 consecutive cases," Radiology **165**, 99-102 (1987).

[4] M. A. Helvie, D. M. Ikeda, and D. D. Adler, "Localization and needle aspiration of breast lesions: complications in 370 cases," AJR **157**, 711-714 (1991).

[5] J. M. Dixon, and T. G. John, "Morbidity after breast biopsy for benign disease in a screened population," Lancet **1**, 128 (1992).

[6] F. M. Hall, J. M. Storella, D. Z. Silverstone, and G. Wyshak, "Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography," Radiology **167**, 353-358 (1988).

[7] D. Cyrlak, "Induced costs of low-cost screening mammography," Radiology **168**, 661-3 (1988).

[8] Y. H. Chang, L. A. Hardesty, C. M. Hakim, T. S. Chang, B. Zheng, W. F. Good, and D. Gur, "Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment," Med Phys **28**, 455-461 (2001).

[9] S. Paquerault, N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Improvement of computerized mass detection on mammograms: Fusion of two-view information," **29**, 238-247 (2002).

[10] J. Y. Lo, J. A. Baker, P. J. Kornguth, and C. E. Floyd, Jr, "Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks," Acad Radiol **6**, 10-15 (1999).

[11] L. Hadjiiski, B. Sahiner, H. Chan, N. Petrick, M. Helvie, and M. Gurcan, "Analysis of temporal changes of mammographic features: computer-aided classification of malignant and benign breast masses," Medical Physics **28**, 2309-17 (2001).

[12] Z. Huo, M. L. Giger, C. J. Vybomy, D. E. Wolverton, and C. E. Metz, "Computerized classification of benign and malignant masses on digitized mammograms: a study of robustness," Academic Radiology **7**, 1077-1084 (2000).

[13] Y. H. Chang, B. Zheng, and D. Gur, "Computer-aided detection of clustered microcalcifications on digitized mammograms: a robustness experiment," Acad Radiol **4**, 415-8 (1997).

[14] G. D. Tourassi, J. Y. Lo, and M. K. Markey, "Validation of a constraint satisfaction neural network for breast cancer diagnosis: new results from 1030 cases," in Medical Imaging 2003: Image Processing, San Diego, CA (San Diego, CA, 2003), p.207-214.

[15] A. O. Bilska-Wolak, and C. E. Floyd Jr, "Tolerance to missing data using a likelihood ratio classifier for computer-aided classification of breast cancer," **submitted**, (2004).

[16] BI-RADS, "American College of Radiology Breast Imaging - Reporting and Data System (BI-RADS) 3rd ed.," American College of Radiology, 1998.

[17] H. L. VanTrees, Detection, Estimation, and Modulation Theory (Part I), (John Wiley & Sons, New York, 1968).

[18] R. N. McDonough, and A. D. Whalen, Detection of Signals in Noise, (Academic Press, San Diego, 1995).

[19] J. P. Egan, Signal detection theory and ROC analysis, (Academic Press, New York, 1975).

[20] C. E. Metz, "Basic principles of ROC analysis," Sem Nuc Med **8**, 283-298 (1978).

[21] C. Metz, "Evaluation of CAD methods," Computer-aided Diagnosis in Medical Imaging, edited by K. Doi et al. (Elsevier Science, Amsterdam, 1998) 543-554.

[22] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," Radiology **201**, 745-750 (1996).

[23] X.-H. Zhou, N. A. Obuchowski, and D. K. McClish, Statistical Methods in Diagnostic Medicine, (John Wiley & Sons, New York, 2002) 437.

[24] E. S. Burnside, D. L. Rubin, R. D. Shachter, R. E. Sohlich, and E. A. Sickles, "A probabilistic expert system that provides automated mammographic–histologic correlation: initial experience," American Journal of Roentgenology **182**, 481- (2004).

[25] A. K. Jain, R. C. Dubes, and C.-C. Chen, "Bootstrap techniques for error estimation," IEEE Trans. Pattern. Anal. Mach. Intell. **9**, 628-633 (1987).

[26] B. Efron, and R. J. Tibshirani, "Improvements on cross-validation: the .632+ bootstrap method," J. Am. Stat. Assoc. **92**, 548-560 (1997).

[27] B. Efron, and R. J. Tibshirani, An Introduction to the Bootstrap, Monographs on Statistics and Applied Probability, ed. D. R. Cox et al. (Chapman & Hall, New York, NY, 1993) 436.

[28] A. O. Bilska-Wolak, C. E. Floyd Jr, L. W. Nolte, and J. Y. Lo, "Application of likelihood ratio to classification of mammographic masses; performance comparison to case-based reasoning.," Medical Physics **30**, 949-958 (2003).

[29] J. Y. Lo, M. K. Markey, J. A. Baker, and C. E. Floyd, Jr, "Cross-institutional evaluation of BI-RADS predictive model for mammographic diagnosis of breast cancer," AJR **178**, 457-463 (2002).